جامعة الأزهر – غزة
كلية الاقتصاد والعلوم الإدارية
قسم الإحصاء



الإحصاء الحيوي

إعداد :
عمر كمال البيروتي
Omar K. Al-beiruty
2009 3844

إشراف الدكتور :
محمود خالد عكاشة
Mahmoud K. Okasha

فهرس المحتويات

الفصل الأول : الإحصاءات الوصفية

الفصل الثاني : المبادئ الأساسية للاستدلال الإحصائي

الفصل الثالث : تحليل البيانات الوصفية

الفصل الرابع : مقارنة المتوسطات

الفصل الخامس : الارتباط والانحدار الخطي البسيط

الفصل السادس : الانحدار الخطي المتعدد

الفصل السابع : تصميم وتحليل التجارب

الفصل الثامن : الانحدار اللوجستي

الفصل التاسع : تحليل البقاء

الفصل الأول

الإحصاءات الوصفية

Descriptive Statistics

- أنواع البيانات
- وصف البيانات الوصفية
- وصف البيانات الكمية
- استكشاف البيانات الكمية

أنواع البيانات

يمكن تصنيف البيانات الإحصائية إلي نوعين رئيسين :

١. البيانات الوصفية (النوعية).
وهي البيانات التي تكون في صورة غير عددية أي لا يمكن قياسها ولها عدد معين من الحالات من الفئات من دون أي وزن لهذه الفئات ومنها على سبيل المثال : لون العينين ، الجنس ، فصيلة الدم ، الديانة . المستوى التعليمي وغيرها وتسمى أيضا البيانات الوصفية بالبيانات الاسمية .

٢. البيانات الكمية (العددية).
هي تلك البيانات التي تكون في صورة عددية أي يمكن قياسها ، ومنها على سبيل المثال : الطول ، الوزن ، الدخل ، عدد الحوادث الشهرية ، عدد أفراد الأسرة ، وتسمى أيضا البيانات الكمية بالبيانات الفئوية .

أولاً : وصف البيانات الوصفية (النوعية)

١- العرض الجدولي للبيانات الوصفية :

وهي عبارة عن وضع البيانات في جدول ويتم ذلك عن طريق تحديد الصفات المختلفة التي تنتمي إليها البيانات وحساب عدد المفردات المناظرة لكل فئة من هذه الصفات ووضع ذلك في جدول .

مثال:
جامعة ما بها 4850 طالب ، موزعين علة 5 كليات مختلفة ، كل كلية تحتوي على عدد من الطلبة ، فكانت كلية العلوم بها 400 طالب وكلية الآداب بها 800 طالب وكلية الحقوق بها 950 طالب وكلية التربية بها 850 طالب وكلية التجارة بها 1350 طالب .
بالتالي تستطيع وضع هذه البيانات بجدول تكراري وحساب نسب أعداد الطلاب في كل كلية كما يلي :

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	الآداب	800	18.4	18.4	18.4
	التجارة	1350	31.0	31.0	49.4
	التربية	850	19.5	19.5	69.0
	الحقوق	950	21.8	21.8	90.8
	العلوم	400	9.2	9.2	100.0
	Total	4350	100.0	100.0	

ويسمى هذا الجدول بالبسيط لأن البيانات تتعلق بظاهرة واحدة أو صفة واحدة فقط .

حيث يستفاد من الجداول التكرارية كونها تسهل فيهم البيانات ، حيث أنها توضح أعداد ونسب أفراد العينة حسب الصفات المشتركة .

٢- العرض البياني للبيانات الوصفية

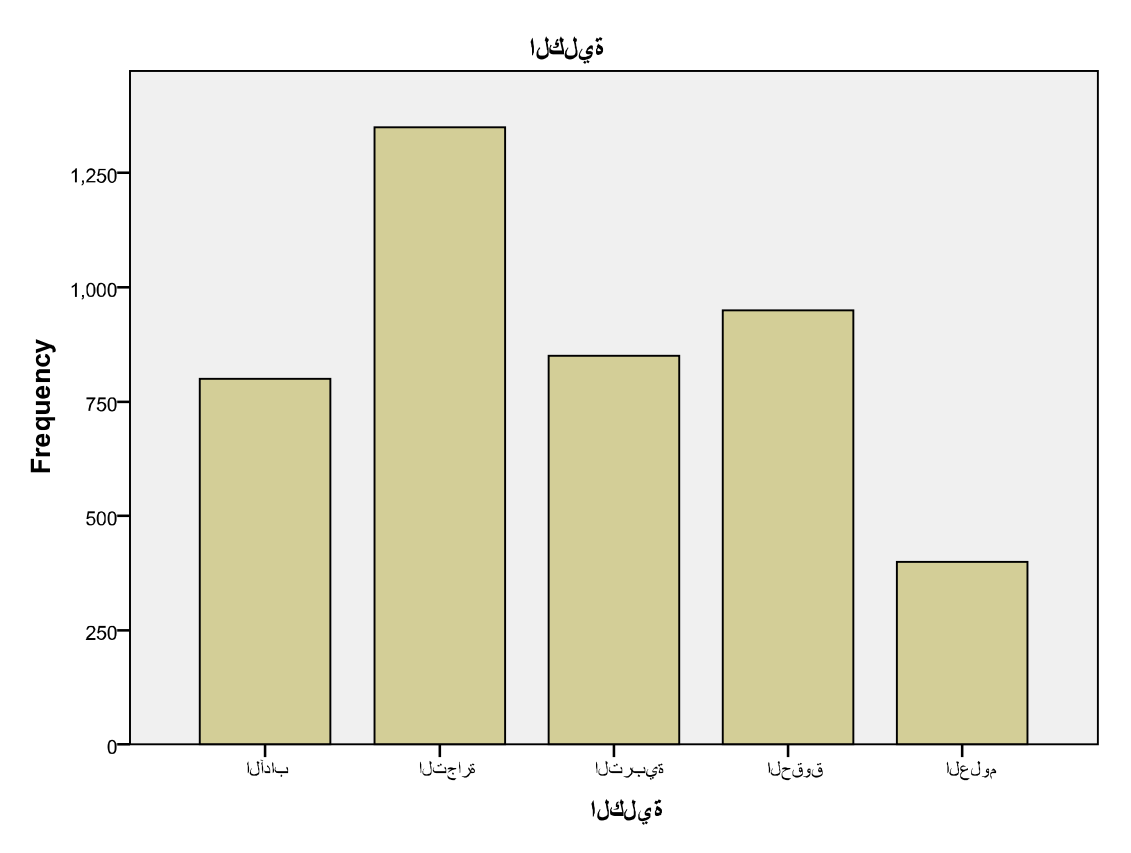
يعتبر استخدام الرسوم البيانية طريقة فعالة في عرض البيانات بشكل واضح يظهر الخصائص الهامة لهذه البيانات وبشكل بسيط وسهل للفهم من خلال النظر إليه .
وتختلف طريقة العرض البياني للبيانات الوصفية من حيث كونها تتعلق بظاهرة واحدة أو ظاهرتين فأكثر .

أ- العرض البياني في حالة ظاهرة واحدة .

من أهم الطرق البيانية التي تستخدم في حالة ظاهرة واحدة هي الأعمدة البسيطة والرسم الدائري .

أ -١- الأعمدة البسيطة

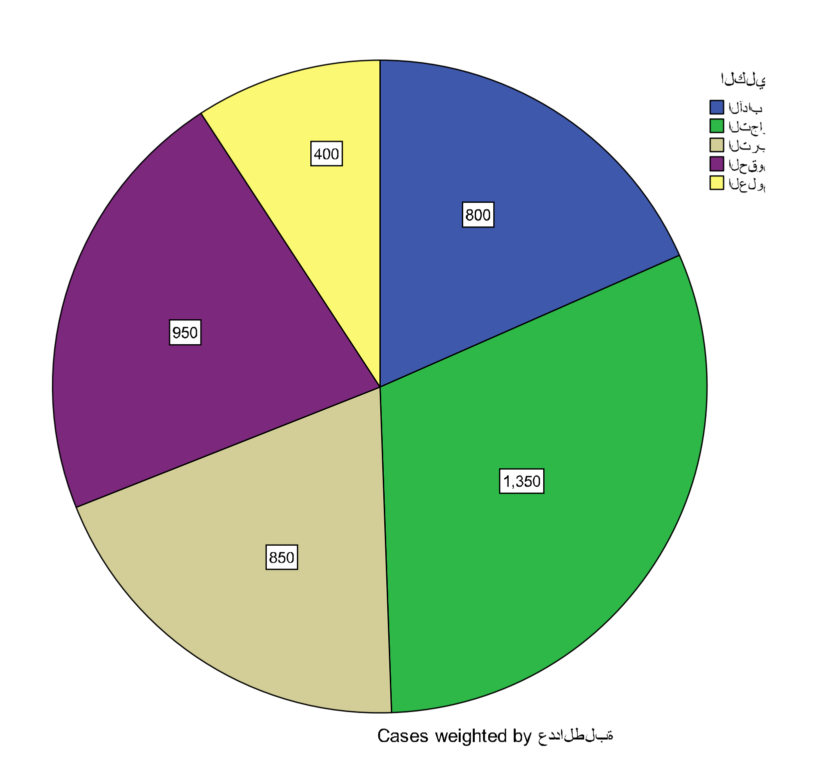
مثال : بافتراض بيانات المثال السابق ونريد تمثيلها بيانياً.



تستخدم طريق الأعمدة البسيطة لمعرفة التطورات والاختلافات الموجودة داخل ظاهرة معينة وتتلخص طريق الأعمدة البسيطة في تمثيل المسميات (الفئات) على المحور الأفقي كما يظهر من المثال السابق ، وقيم هذه المسميات على المحور العمودي ، بحيث يتم رسم مستطيل على كل مسمى (فئة) ويكون ارتفاعها يمثل القيمة التي تقابل ذلك المسمى وذلك باستخدام مقياس رسم مناسب .

أ – ٢ - الرسم بالدائرة :

وتعتبر هذه الطريق من أفضل الطرق لتمثيل البيانات ذات الصفة المشتركة ونستطيع من خلالها أن نقارن الأجزاء مع بعضها البعض .



بمجرد النظر إلي رسم الدائرة نستطيع التعرف على البيانات من حيث التركيز ، ومن الصفات التي تستحوذ على أكبر قيمة وغيرها من خصائص البيانات ، بسهولة .

ب - العرض البياني في حالة ظاهرتين فأكثر :

من أهم الطرق البيانية التي تستخدم في حالة ظاهرتين فأكثر الأعمدة المتلاصقة (المزدوجة) والأعمدة المجزأة .

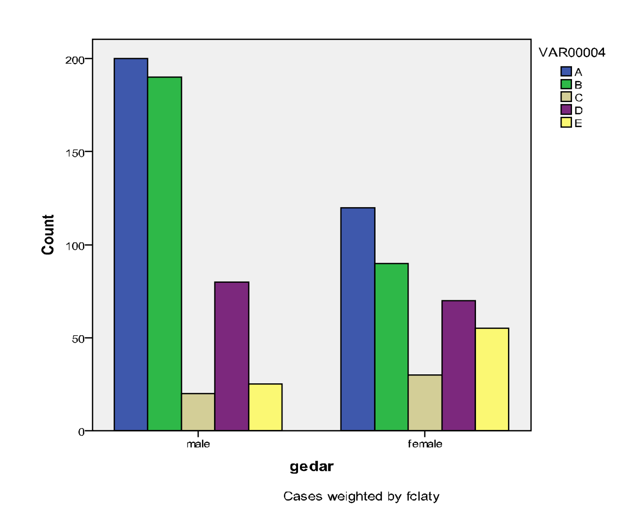
ب – ٢ - الأعمدة المتلاصقة (المجزاة):

تستخدم هذه الطريقة عندما يكون المطلوب هو المقارنة بين ظاهرتين فأكثر .

مثال :

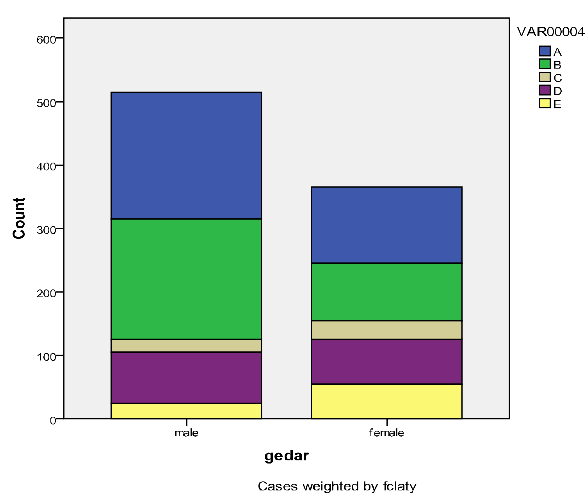
البيانات التالية تمثل أعداد الطلبة في أقسام كلية التجارة في جامعة ما حسب الجنس .

القسم	طلاب	طالبات
الإدارة	200	120
المحاسبة	190	90
الإحصاء	20	30
العلوم السياسة	80	70
الاقتصاد	25	55



ب – ٣ - الأعمدة المجزأة :

تستخدم هذه الطريقة أيضا للمقارنة بين الظواهر ، وذلك برسم عمود واحد يمثل كل الظاهرة المراد دراستها ، ثم تقسم كل عمود لعدة أجزاء وهذه الأجزاء هي أقسام الظاهرة أو مسميات الظاهرة بحيث يتناسب كل جزء مع قيمته .



ثانيا : وصف البيانات الكمية

هناك عدة مقاييس إحصائية لوصف البيانات الكمية منها :

i. مقاييس النزعة المركزية (مقاييس الموضع) ومنها :

- **الوسط الحسابي**
- **الوسيط**
- **المنوال**

ii. مقاييس التشتت ومنها :

- **المدى**
- **نصف المدى الربيعي (الانحراف الربيعي)**
- **التباين والانحراف المعياري**

iii. مقاييس التشتت النسبي ومنها :

- **معامل الاختلاف**

أولاً : مقاييس النزعة المركزية (مقاييس الموضع)

١. الوسط الحسابي

يعرف الوسط الحسابي لمجموعة من المشاهدات أنه مجموع هذه المشاهدات (القيم) مقسوما على عدد المشاهدات (القيم) ، ويرمز له بالرمز $\overline{x}$.

مثال :لحساب الوسط الحسابي لمجوعة القيم التالية (77,69,91,73,87)

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}, \qquad = \frac{77+69+91+73+87}{5} = 79.4$$

من خواص الوسط الحسابي أن مجموع انحرافات القيم عن وسطها الحسابي تساوي صفر دائما .

٢. الوسيط

يعرف الوسيط لمجموع من البيانات بأنه القيمة التي تتوسط تلك البيانات بعد ترتبيها تصاعدياً أو تنازلياً ، أي هو القيم التي تقسم البيانات إلى قسمين متساويين في العدد ويرمز له بالرمز M_e .

٣. المنوال

يعرف المنوال لمجموعة من البيانات بأنه القيمة الأكثر تكراراً أو شيوعاً ، وعليه فإن القيمة التي تتكرر أكثر من بقية القيم تعرف بأنها هي المنوال .

ثانيا : مقاييس التشتت :

تعتبر المتوسطات أحد المقاييس الوصفة الإحصائية التي نحتاجها لوصف البيانات وصفا كمياً ، إلا أن إيجاد احد المتوسطات لا يعطي وصفا كاملا للبيانات لأنة بين القيم التي تتركز عندها البيانات دون أن يعطي أي فكرة عن مدي تقارب أو تباعد البيانات ن بعضها البعض أو عن قيمة المتوسط ذاتها ، أي أن قيمه المتوسط لا تبن مدي تجانس البيانات أو تشتتها ، ولذلك فإننا نحتاج إلي مقاييس أخرى تبين مدى تشتت البيانات وهذه المقاييس تمسى بمقاييس التشتت ، ومن أهمها :

١. المدى
٢. نصف المدى الربيعي (الانحراف الربيعي)
٣. التباين والانحراف المعياري

١. المدى

المدى هو من ابسط مقاييس التشتت ويعرف بأنه الفرق بين أكبر و اصغر قيمة في مجموعة البيانات ويرمز له بالرمز R أي أن :

$$R = X\,max - X\,min$$

مثال : احسب المدى للبيانات التالية :

أ. 69 , 92 , 73 , 78 , 65 ب. 50 , 35 , 42 , 26 , 33 , 25

الحل (أ) : المدى = أكبر قيمة – أصغر قيمة = 92 - 7 = 85

الحل (ب) : المدى = أكبر قيمة – أصغر قيمة = 50 - 25 = 25

نلاحظ من السابق أن المجموع (أ) اكبر تشتتا من المجموعة (ب) لان قيمة المدى اكبر .

٢ . نصف المدى الربيعي (الانحراف الربيعي)

يعاب على المدى أنه يتأثر بالقيم المتطرفة وللتخلص من هذا العيب ممكن أن نستخدم طريقة حساب نصف المدى الربيع وذلك باستبعاد الربع الأول والربع الأخير من القيم ويحسب المدى للقيم المتبقية

$$Q = \frac{Q_3 - Q_1}{2}$$

٣ . التباين والانحراف المعياري

يعرف تباين مجموع من القيم بأنه متوسط مجموع مربعات القيم عن وسطها الحسابي ، وبالتالي فإن وحدات التباين هي مربع وحدات القيم الأصلية .

والانحراف المعياري لمجموعة من البيانات هو الجذر التربيع الموجب للتباين وبالتالي فان وحدات الانحراف المعياري هي نفس وحدات البيانات الأصلية ويرمز له بالرمز S .

قانون حساب التباين

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}.$$

قانون حساب الانحراف المعياري

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

ومن مزايا التباين والانحراف المعياري كونه من أهم المقاييس في تعيين درجة التشتت ، أيضا يتناول جميع القيم للبيانات ، أيضا كونه أدق مقاييس التشتت .

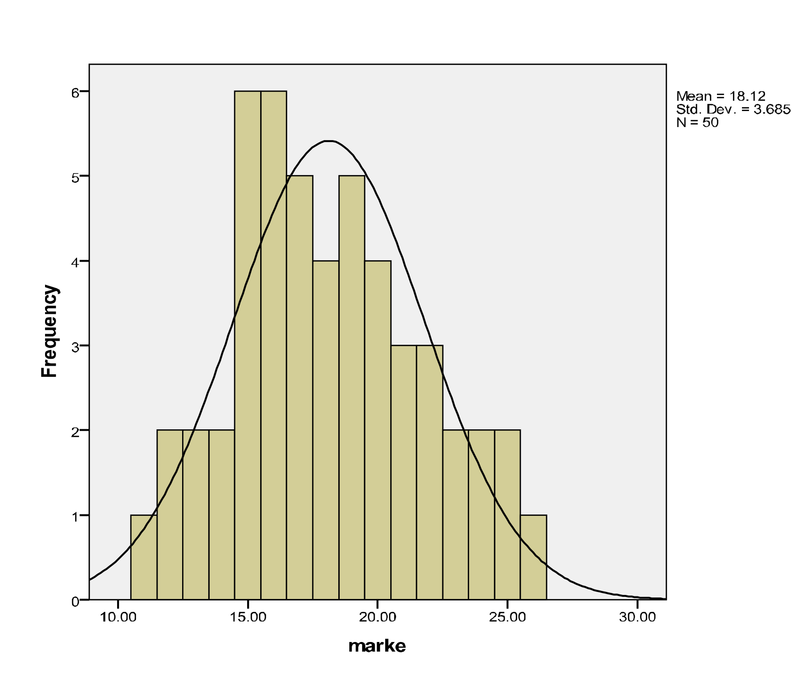
عرض (استكشاف) المتغيرات الكمية

اختبار طبيعية البيانات

يمكن اختبار التوزيع الطبيعي للبيانات عن طريق رسم المدرج التكراري مع المنحنى الطبيعي في رسم بياني واحد، فإذا كانت غالبية أعمدة المدرج التكراري تقع تحت المنحنى حينها يمكن القول بان البيانات تتبع التوزيع الطبيعي ، أيضا يمكن اكتشاف ما إذا كانت البيانات بها التواء أم لا ، طريق ثانية لاستكشاف توزع البيانات ، عن طريق رسم الصندوق Box Plot ، عن طريقه أيضا يمكن استكشاف البيانات ومعرفة ماذا إذا كانت بها قيم شاذة أو قيم متطرفة أو كلاهما ، أيضا طريقة ثالثة باستخدام رسم Q-Q Plot ، أيضا رسم شكل الانتشار مع خط 45 درجة مئوية ، وأخير القول النهائي لطبيعية البيانات يقرر من الاختبار الإحصائي كلمنجروف سمير نوف ، وسوف نتعرض لهم بشيء من بالتفصيل .

١. رسم المدرج التكراري مع المنحنى الطبيعي .

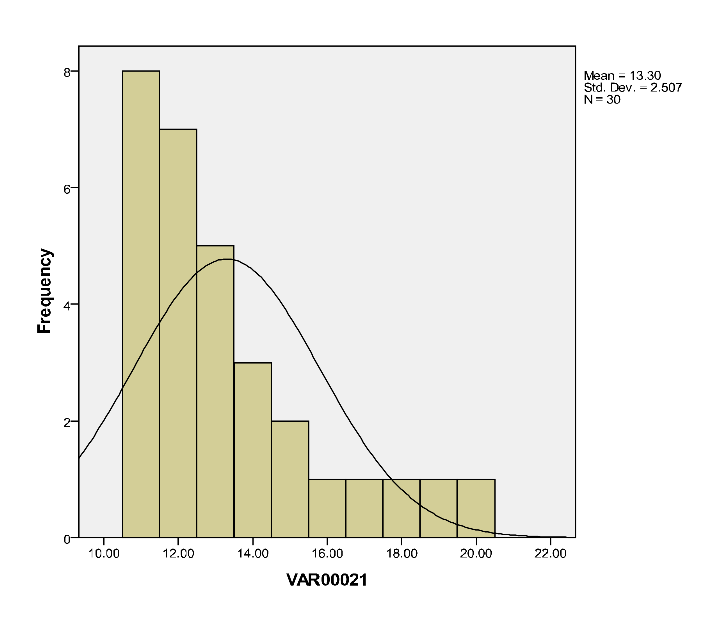
مثال لدينا بيانات عن علامات 50 طالب في امتحان ما كالتالي
٢١,١٥,١٥,١٣,١٦,١٥,١٧,١٩,٢٢,٢١,١٨,١٥,١٩,٢٠,١٦,١٥,١٦ ٢١
١٧,١٧,١٨,١٨,١٩١٦,٢٠,٢٥,٢٦,٢٣,١٨,١٧,١٦,١٤,١٥,١٣,١٢,٢٠,٢٢,١٩,١٢,١١,١٧,١٦,٢٠,٢٤,٢٥,٢٣,٢٢,٢٤
١٤,١٩



حيث نلاحظ من الشكل السابق بان غالبية الأعمدة تقع تحت المنحنى ، مما يعطي دلالة واضحة على أن البيانات تتبع التوزيع الطبيعي.

مثال 2 : السم البياني التالي يمثل مبيعات الشركة في شهر نوفمبر .

حيث نلاحظ من الرسم بأن البيانات ملتوية نحو اليمين ، مما يعني بأن مبيعات الشركة لا تتبع التوزيع الطبيعي .

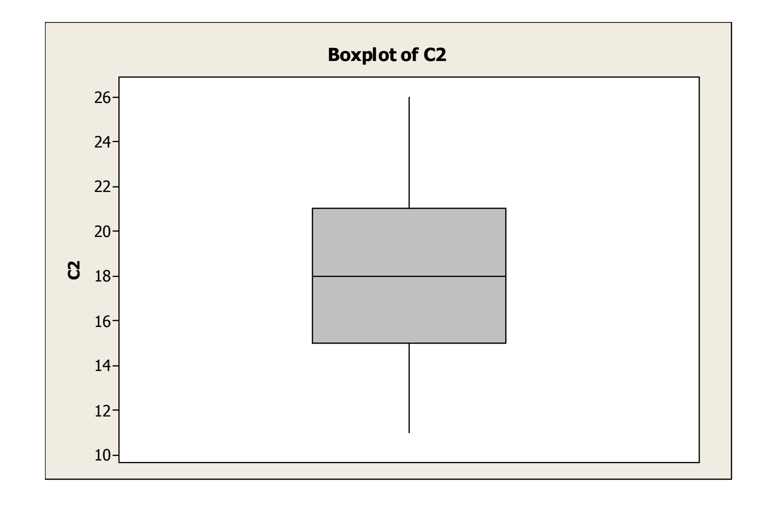


٢. رسم الصناديق Box Plot

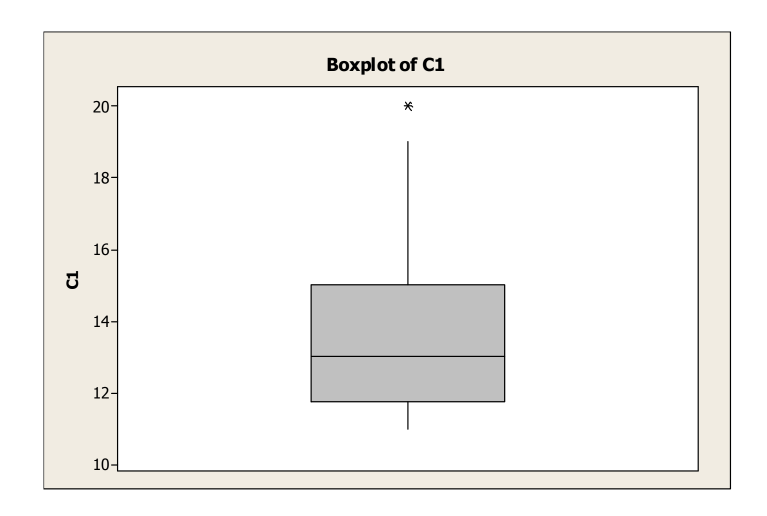
رسم الصندوق هو شكل بياني يوضح المدى الذي تنتشر علية البيانات ونمط الاختلاف ودرجة التماثل والالتواء في توزيع البيانات وكذلك في توزيع النصف الأوسط للبيانات .

مثال

بالتطبيق على بيانات مثال رقم واحد وبرسم الصندوق يظهر لدينا الرسم التالي :



أما بالتطبيق على بيانات مثال ٢ :



يوضح الرسم البياني رسم الصندوق والشعيرات لتوزيع بيانات بها التواء نحو اليمين لأن الشعيرة اليمنى أطول من الشعيرة اليسرى ، كما يوضح الشكل وجود قيمة شاذة .

٣. اختبار كلمنجروف سمير نوف

بالتطبيق على بيانات مثال رقم واحدا ، وبعد إدخال البيانات في برنامج التحليل الإحصائي SPSS وطلب أمر اختبار التوزيع الطبيعي كلمنجروف سمير نوف ينتج لدينا .

One-Sample Kolmogorov-Smirnov Test

		العمر بالسنوات
N		97
Normal Parameters[a,b]	Mean	39.99
	Std. Deviation	10.251
Most Extreme Differences	Absolute	.087
	Positive	.087
	Negative	-.051-
Kolmogorov-Smirnov Z		.859
Asymp. Sig. (2-tailed)		.452

a. Test distribution is Normal.

b. Calculated from data.

وبما أن قيمة الـ Sig أكبر من P-value 0.05 مما يجل على عدم معنوية الاختبار أي أن علامات الطلاب تتبع التوزيع الطبيعي.

الفصل الثاني

المبادئ الأساسية للاستدلال الإحصائي

Basic Principles of Statistical Inference

- تقدير المعالم
- التقدير بنقطة والتقدير بفترة ثقة
- اختبار الفرضيات الإحصائية
- الأخطاء في اختبار الفرضيات
- اختبار الفرضيات حول المتوسط
- اختبار الفرضيات حول الفرق بين متوسطين
- اختبار الفرضيات حول تباين مجتمع واحد

تقدير المعالم

أحد المشاكل الهامة في الاستدلال الإحصائي هي مشكلة تقدير معالم المجتمع المجهولة مثل (متوسط المجتمع ، تباين المجتمع ، نسبة الحدوث في المجتمع ، الفرق بين متوسطي مجتمعين ...) فإننا نواجه باحتمال الوقوع في الخطأ عند التقدير، ويجب تحديد حجم هذه الخطأ لتظهر مدى الدقة في التقديرات المبنية على نتائج العينة .

التقدير بنقطة والتقدير بفترة ثقة :

إذا قدرت معلمة المجتمع بقيمة وحيدة فهذا يسمى بتقدير المعلمة بنقطة ، وهذا التقدير قد يكون قريبا جدا من المعلمة المجهولة، ولكن غالبا لا يطابق القيمة الفعلية لهذه المعلمة ولكن إذا انتهت عملية التقدير عند هذا الحد فإننا لن نعلم مدى دقة التقدير أو مدى بعده من القيمة الحقيقة المجهولة ، أما إذا حاولنا تقدير معلمة المجتمع قيد الدراسة بقيمتين ،بحيث يمكن اعتبار أن المعلمة تقع بينهما فإننا نحصل على ما يسمى بالتقدير بفترة ثقة لهذه المعلمة

فترة ثقة للوسط الحسابي

مثال

اشترت أحدى الشركات ماكينة لتعبئة أكياس الزر اتوماتيكيا بحيث يكون وزن الكيس5 كجم ولاختبار هذه الماكينة تم أخذ عينة من 30 كيس وتم وزنهم على ميزان دقيق ، فكان متوسط وزن الكيس في العينة هو 0.2267 كجم بانحراف معياري للعينة 0.5252 كجم ، أوجد 95٪ فترة ثقة لمتوسط أوزان أكياس الرز التي تقوم الماكينة بتعبئتها .

الحل

$$\bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}},$$

$$0.2267 \pm 1.96 \times \frac{0.5252}{\sqrt{30}},$$

$$(0.0388, 0.4146).$$

التفسير :

أنه إذا تم اخذ عينات كثيرة جدا وكل منها بحجم 30 كيس فننا سوف نجد أن 95٪ من هذه العينات تعطي متوسط حسابي لوزن الكيس يقع في فترة الثقة 0.0388 و 0.4146 كجم، وأن 5٪ فقط من هذه العينات سوف يكون متوسطها خارج هذه الفترة .

القريب لتوزيع t :

أما عندما تكون قيمة تباين المجتمع σ^2 مجهولة وفي نفس الوقت حجم العينة صغير $(n<30)$ فانه يمكن التعبير عن توزيع المعاينة للوسط الحسابي بالمقدار : $T = (\bar{X} - \mu)/(s/\sqrt{n})$ الذي يتبع توزيع t بدرجات حرية n-1

وبالتالي تكون فترة ثقة للوسط الحسابي للمجتمع μ هي :

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}},$$

مثال

يملك السيد حسن محطة لتعبئة الوقود في إحدى المدن ، وقام باختيار عينة من الزبائن مكونة من ١٤ أشخاص قاموا بتعبئة سياراتهم بالبنزين ، وتم تسجيل كميات البنزين المشتراه ، فوجد أن هذه العينة قد أعطت متوسطا قدرة ٤١٢ لترا وانحراف معياري ٢,٣١ لترا والمطلوب تقدير ٩٥٪ فترة ثقة لمتوسط كميات البنزين المشتراه بواسطة الزبون الواحد من هذه المحطة .

$$\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}},$$

$$412 \pm 2.1604 \frac{231}{\sqrt{14}},$$

$$(278.6, 545.4),$$

فترة ثقة للنسبة

إذا كان حجم العينة كبيرا فانه يمكن استنتاج أن النسبة المحسوبة من بيانات عينة كبيرة P تتبع التوزيع الطبيعي بتوقع π (النسبة المجهولة في المجتمع) ، وتباين $\pi(1-\pi)/n$ ، أي انه يمكن صياغة توزيع المعاينة لنسبة الحدوث في العينة على الصورة :

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

وهذه الدالة تتبع التوزيع الطبيعي المعياري.

مثال

خلا سنة ١٩٨٨ تم إجراء استطلاع لعينة من ٤٠٠ موظف حول رأيهم بقضية ما ، فأجاب ٣٢٠ منهم بالموافقة ، كون ٩٥٪ فترة ثقة لنسبة المؤيدين للقضية .

الحل

أولا : نحسب نسبة الحدوث في العينة

$$P = 320 / 400 = 0.80$$

وحيث أن حجم العينة كبيرا فيمكن تقريب توزيع المعاينة للنسبة إلي التوزيع الطبيعي

$$Z = \frac{P - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

ومنها نجد أن فترة الثقة المطلوبة هي :

$$P \pm Z_{\alpha/2}\sqrt{\frac{P(1-p)}{n}}$$

$$0.80 \pm 1.96\sqrt{\frac{0.80*0.20}{400}}$$

$$0.80 \pm 0.0392$$

$$(0.7608 , 0.8392)$$

اختبار الفرضيات الإحصائية Hypothesis Testing

نظرا لعدم معرفة قيم المجمع في الحياة العملية ، فان القيم المحسوبة من بيانات العينة تستخدم ليس فقط لتقدير هذه المعالم ب أيضا للحكم على مدى صحة أو خطأ رأي معين متعلق بهذه المعالم .

أن أي تعبير يتعلق بالمعلمة المجهولة في المجتمع يسمى فرضية إحصائية حول هذه المعلمة ، وكذلك فان التعبير الذي يتناقض مع التعبير الأول حول المعلمة المجهولة يسمى أيضا فرضية إحصائية ، ولذلك فان هناك فرضيتين إحصائيتين متضادتين يطلق علي أحدهما فرضية عدمية والأخرى فرضية بديلة ، وأن محاولة التحقق من صحة أي من هذه الفرضيات ما هي إلا محاولة لاختبار الفرضية العدمية واتخاذ قرار برفض أو قبول هذه الفرضية يتم بناءا على المعلومات البسيطة التي نحصل عليها من بيانات العينة (التي قد تكون صغيرة) وهذا ما يسمى باختبار الفرضية الإحصائية

الأخطاء في اختبار الفرضيات

- **الخطأ من النوع الأول** :هو احتمال رفض الفرضية العديمة علما بأنها صحيحة
- **الخطأ من النوع الثاني** : وهو احتمال قبول الفرضية العديمة علما بأنها خاطئة

مستوى المعنوية وقيمة P-value وقوة الاختبار :

- **مستوى المعنوية :** وهي قبول الفرضية العديمة علما بأنها صحيحة .

- P-value: وهي احتمال الحصول على قيمة أكثر تطرفا من القيمة المحسوبة.

- أما **قوة الاحتيار الإحصائي** : وهِي عبارة عن احتمال رفض الفرضية العديمة عندما تكون خاطئة فعلا.

اختبار الفرضيات حول المتوسط

يمكن صياغة الفرضية العدمية كما يلي والفرضية البديلة تأخذ أحد أشكال الحالات التالية :

$$H_0: \mu = \mu_0$$
$$H_a: \mu \neq \mu_0$$
$$\mu < \mu_0$$
$$\mu > \mu_0$$

وكذلك فان قيمة دالة الاختبار سوف تختلف تبعا لاختلاف الحالات التالية :

١. عندما يكون حجم العينة كبيرا و σ^2 مجهولة :

في هذه الحالة فان دالة الاختبار تكون على الصورة

$$Z = \frac{\ddot{X} - \mu}{S/\sqrt{n}}$$

٢. عندما يكون حجم العينة كبيرا و σ^2 معلومة :

في هذه الحالة فان دالة الاختبار تكون على الصورة

$$Z = \frac{\ddot{X} - \mu}{\sigma/\sqrt{n}}$$

٣. عندما يكون حجم العينة صغيراً و σ^2 مجهولة :

في هذه الحالة فان دالة الاختبار تكون على الصورة

$$t_{n-t} = \frac{\ddot{X} - \mu}{S/\sqrt{n}}$$

مثال

اختيرت عينة من ٢٠٠ شخص بطريق عشوائية ،من سجلات الوفيات بإحدى المناطق فوجد أن الوسط الحسابي والانحراف المعياري لعمر المتوفى هما ٦١ و ١٠ على التوالي ، أختبر الفرضية القائلة بأن متوسط عدد سنوات بقاء الشخص على قيد الحياة في تلك المنطقة يساوي ٦٠ سنة عند مستوى معنوية ٠,٠٥ .

الحل

لدينا $n=200$, $\ddot{X}=60$, $S = 10$, $\mu_0= 60$, $\alpha = 0.05$

وتكون الفرضية العدمية والفرضية البديلة كالآتي :

$H_0: \mu = 60$

$H_a: \mu \neq 60$

ونظرا لان حجم العينة كبيرا فان دالة الاختبار :

$$Z = \frac{\ddot{X} - \mu}{S/\sqrt{n}}$$

وتكون قيمة Z المحسوبة من بيانات العينة كما يلي :

$$Z =(61-60) / (10/\sqrt{200})$$

$$= 1.414$$

وحيث أن هذه القمة لا تقع في منطقة رفض الفرض العدمي ، فانه لا يوجد دليل كافي من بيانات العينة يمكننا من رفض الفرضية العدمية . أي نقبل الفرض القائل بأن متوسط بقاء الشخص على قيد الحياة يساوي 60 سنة عندي مستوى معنوية 0.05 .

اختبار الفرضيات حول الفرق بين متوسطين حسابيين من عينتين مستقلتين

هناك عدة حالات تختلف فيها دالة الاختبار وكذلك التوزيع الاحتمالي للدالة الاختبار تبعا لحجم العينتين والمعلومات المتوفرة عن تبايني المجتمعين ، وهذه الحالات الممكنة هي :

١. عندما يكون حجمي العينتين كبير

$$Z = \frac{(\ddot{X}_1 - \ddot{X}_2) - d}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}}$$

٢. عندما تكون قيمتي تبايني المجتمعين σ_1^2 و σ_2^2 معلومتين

$$Z = \frac{(\ddot{X}_1 - \ddot{X}_2) - d}{\sqrt{\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}}}$$

٣. عندما يكون تباينا المجتمعين σ_1^2 و σ_2^2 مجهولين ولكن معلوم أنها متساويين في القيمة وكانت العينتين صغيرتي الحجم

$$S^2{}_p = \frac{(n_1 - 1)\, S^2_1 + (n_2 - 1)\, S^2_2}{n1 + n2 - 2}$$

وفي هذه الحالة فان دالة الاختبار سوف تأخذ الصورة

$$T_{n1+n2-2} = \frac{(\ddot{X}_1 - \ddot{X}_2) - d}{S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

٤. عندما يكون تباين المجتمعين σ_1^2 و σ_2^2 مجهولين تماما وغير متساويين والعينتين صغيرتي الحجم

$$t_f = \frac{(\ddot{X}_1 - \ddot{X}_2) - d}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

وهي تتبع توزيع t بدرجات حرية f .

مثال

أخذت عينة عشوائية مكونة من 80 أسرة مدينة أ فكان الوسط الحسابي والانحراف المعياري للدخل الشهري للأسرة المحسوبين من بيانات العينة هما 250 و 40 دينارا على الترتيب ، وأخذت عينة عشوائية أخرى مكونة من 100 أسرة من مدينة ب فكان الوسط الحسابي والانحراف المعياري المحسوبين من بيانات العينة 270 و 45 دينارا على الترتيب ، أختبر الفرضية القائلة بتساوي متوسطي الدخل الشهري للأسرة في المدينتين أ و ب عند مستوى معنوية 0.01

الحل

من البيانات المعطاة في هذه المثال نلاحظ أن :

$n_1 = 80$, $\ddot{X}_1 = 250$, $S_1 = 40$, $n_2 = 100$, $\ddot{X}_2 = 270$, $S_2 = 45$
$\alpha = 0.01$, $d = 0$

وتكون الفرضية العديمة والفرضية البديلة كالتالي :

$H_0: \mu_1 - \mu_2 = 0$
$H_a: \mu_1 - \mu_2 \neq 0$

وحيث أن حجم العينتين كبير فانه يمكن تقدير تبايني المجتمعين σ_1^2 و σ_2^2 من العينية باستخدام S_1^2 و S_2^2 اللتان يمكن استخدامهما في دالة الاختبار فيكون :

$$Z = \frac{(\ddot{X}_1 - \ddot{X}_2) - d}{\sqrt{\frac{S^2_1}{n_1} + \frac{S^2_2}{n_2}}}$$

$$Z = \frac{(250 - 270) - 0}{\sqrt{\frac{1600}{80} + \frac{2025}{100}}}$$

$$= -20 / 6.344 = - 3.152$$

وحيث أن القيمة المطلقة ل Z المحسوبة من بيانات العينة أكبر من Z المستخرجة من الجدول وهي القيمة الحرجة ، وتقع تلك القيمة في منطقة رفض الفرض العدمي ، فإننا نرفض الفرض العدمي ونقبل الفرضية البديلة بمستوى معنوية 0.01 ، أي أن هناك دليل كافي على وجود فرق حقيقي بين متوسطي الدخل الشهري للأسرة في المدينتين أ و ب .

اختبار الفرضيات حول تباين مجتمع واحد

في هذه الحالة يكون الاهتمام منصبا على اختبار فرضيات حول التباين كمقياس تشتت لمجتمع واحد .
وتكون الفرضية العدمية في هذه الحالة هي :

$$H_0: \sigma^2 = \sigma_0^2$$

والفرضية البديلة

$$H_0: \sigma^2 \neq \sigma_0^2$$
$$\sigma^2 < \sigma_0^2$$
$$\sigma^2 > \sigma_0^2$$

حيث σ^2 هي تباين المجتمع المجهول ، بينما σ_0^2 هي قيمة محددة افتراضية.
ولاختبار صحة الفرضية العدمية نستخدم دالة الاختبار :

$$\chi^2_{(n-1)} \quad \frac{(n-1)s^2}{\sigma^2}$$

مثال :

أخذت عينة عشوائية مكونة من ١٦ طالب وسجلت درجاتهم في أحد الامتحانات ، فوجد أن الانحراف المعياري المحسوب من العينة هو ٧ درجات فإذا كان σ^2 ترمز لتباين درجات الطلاب في هذه الامتحان فالمطلوب اختبار الفرضية العدمية التالية ، عند مستوى معنوية ٠,٠١ .

$H_0: \sigma^2 = 36$
$H_0: \sigma^2 > 36$

الحل

من البيانات المعطاة في هذا المثال نلاحظ أن :

$n=16$ $S^2=16$ $\sigma_0^2=36$ $\alpha = 0.01$

وان دالة الاختبار للفروض هي

$$\chi^2_{(n-1)} \qquad \frac{(n-1)s^2}{\sigma^2} \qquad = (15 * 49)/36 = 20.417$$

وهذه تتبع توزيع X^2 بدرجات حرية 15=1-16.

وحيث أن X^2 المحسوبة من بيانات العينة 20.417 اصغر من القيمة الحرجة 30.578 أي تقع في منطقة القبول ، فانه لا يوجد دليل كافي لرفض الفرضية العدمية عند مستوى معنوية 0.01 .

اختبار الفرضيات حول تساوي تبايني مجتمعين مستقلين :

إن الإحصاء المناسب لاختبار تساوي تبايني مجتمعين أي تجانس مجتمعين ، في مستنبط من التوزيع الاحتمالي للنسبة بين تباينين ، ولذلك فان اختبار تجانس مجتمعين المشار إليه يمكن أن يتم باختبار الفرضية العدمية :

$H_0: \sigma_1^2 = \sigma_2^2$

$H_0: \sigma_1^2 \neq \sigma_2^2$ والفرضية البديلة

$\sigma_1^2 < \sigma_2^2$

$\sigma_1^2 > \sigma_2^2$

ولهذا فان هذا الاختبار يسمى أحيانا باختبار النسبة بين تباينين ، وتنتج دالة الاختبار ببساطة بقسمة التباينين المحسوبين من بيانات العينتين المحسوبتين ، أي أن دالة الاختبار في هذا الحالة هي :

$$F_{(n1-1,n2-1)} = \frac{S_1^2}{S_2^2}$$

مثال :

أخذت عينة عشوائية مكونة من ١٦ من الطابعين الرجال وعينة عشوائية أخرى مكونة من ١٢ من الطابعات النساء ، وسجل الوقت اللازم لطابعة رسالة معينة لكل شخص في العينتين ،فوجد أن تباين الوقت اللازم لطباعة هذه الرسالة المحسوب من عينة الرجال يساوي $S_1^2 = 36$ ، والمحسوب من عينة النساء يساوي $S_2^2 = 27$ ، والمطلوب اختبار الفرضية العدمية التالية :

$H_0: \sigma_1^2 = \sigma_2^2$

$H_0: \sigma_1^2 > \sigma_2^2$ والفرضية البديلة

عند مستوى معنوية 0.05 وذلك بافتراض أن الوقت اللازم لطباعة الرسالة لكن من الرجال والنساء يتبع التوزيع الطبيعي .

الحل

دالة الاختبار المحسوبة من بيانات العينتين هي :

$$F_{(15,11)} = 36/27 = 1.333$$

وحيث أن F المحسوبة من بيانات العينة أصغر من F المستخرجة من الجدول ، أي تقع في منطقة القبول فانه لا يوجد دليل كافي لرفض الفرضية العدمية عند مستوى معنوية 0.05 .

الفصل الثالث

تحليل البيانات الوصفية

Statistical Inference on Categorical Variables

- وصف البيانات الوصفية كنسب
- التوزيعات الإحصائية للبيانات الوصفية
- تقريب توزيع ذات الحدين إلي التوزيع الطبيعي
- اختبار الفرضيات حول نسبة الحدوث في المجتمع

البيانات الوصفية .

هي البيانات التي تكون في صورة غير عددية أي لا يمكن قياسها ولها عدد معين من الحالات من الفئات من دون أي وزن لهذه الفئات ومنها على سبيل المثال : لون العينين ، الجنس ، فصيلة الدم ، الديانة . المستوى التعليمي وغيرها وتسمى أيضا البيانات الوصفية بالبيانات الاسمية .

وصف البيانات الوصفية كنسب

مثال

لنفرض أن توزيع طلاب كلية الاقتصاد والعلوم الإدارية على الأقسام المختلفة كان كالتالي :

١٢٠ طالب في قسم الإدارة ، ١٠٠ طالب في قسم المحاسبة ، ٦٠ طالب في قسم العلوم السياسة ، ٥٠ في قسم الاقتصاد ، ٢٠ طالب في قسم الإحصاء ، عندها يمكننا حسابي نسبة الطلاب حسب توزيعهم في كل قسم كما يلي

نسبة طلاب قسم الإدارة = ١٢٠ / ٣٥٠ = ٠,٣٤ ٪
نسبة طلاب قسم المحاسبة = ١٠٠ / ٣٥٠ = ٠,٢٩ ٪
نسبة طلاب قسم العلوم السياسية = ٦٠ / ٣٥٠ = ٠,١٧ ٪
نسبة طلاب قسم الاقتصاد = ٥٠ /٣٥٠ = ٠,١٤ ٪
نسبة طلاب قسم الإحصاء = ٢٠ / ٣٥٠ = ٠,٠٦ ٪

التوزيعات الإحصائية للبيانات الوصفية

توزيع ذات الحدين

هو توزيع احتمالي لمتغير عشوائي متقطع ، وينتج عن تجربة عشوائية تحتمل نتيجتين فقط هما إما نجاح باحتمال P وإما فشل باحتمال $1-P$ ، ويتم تكرار التجربة عدد ثابت n من المرات ويكون المتغير العشوائي X هو عدد مرات ظهور النجاح في تلك التجربة ، وبالتالي فان القيم التي يأخذها المتغير العشوائي X هي : $X = 1,2,...,n$ ويكون التوزيع الاحتمالي لهذا المتغير هو :

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

مثال : في عائلة مكون من ٥ أطفال ،أحسب احتمال أن يكون من بينهم ٣ ذكور علما بأن نسبة الذكور إلي الإناث هي ١:١ .

$n=5$, $p = 0.5$, $q = 0.5$

$P (X = 3) = (5 \;\; 3) \; (0.5)^3 (0.5)^2 = 0.3125$

تقريب توزيع ذات الحدين إلي التوزيع الطبيعي

يمكن تقريب توزيع ذات الحدين إلي التوزيع الطبيعي باستخدام التحويلة التالية :

$$P(X=x)=P\left(\frac{X-np}{\sqrt{np(1-p)}}=\frac{x-np}{\sqrt{np(1-p)}}\right)=P(Z=z)$$

مثال :

بفرض أن احتمال شخص أن يصيب الهدف هو 0.3 ، وبفرض انه تم التصويب على الهدف 100مرة ، احسب احتمال أن هذا الشخص سوف يصيب الهدف 20 و 40.

إذن متوسط إصابة الهدف يساوي $0.3 * 100 = 30$ ، والانحراف المعياري يساوي

$$\sqrt{np(1-p)}=\sqrt{100(0.3)(0.7)}=4.58.$$

$$P(20\le X\le 40)=P\left(\frac{20-0.5-30}{4.58}\le\frac{X-30}{4.58}\le\frac{40+0.5-30}{4.58}\right)$$
$$=P(-2.29\le z\le 2.29)$$
$$=0.9781$$

ولو استخدمنا قانون ذات الحدين نستنتج نفس الاحتمال

$$P(20\le X\le 40)=\sum_{x=20}^{x=40}\binom{100}{x}(0.3)^x(1-0.3)^{100-x},$$

we get a probability of 0.9786.

اختبار الفرضيات حول نسبة الحدوث في المجتمع

في هذه الحالة يكون اهتمامنا منصبا حول اختبار الفروض المتعلقة بنسبة الحدوث في المجتمع، وبالتحديد يهمنا اختبار ما إذا كانت نسبة وقوع حدث معين في مجتمع ما والتي عادة ما تكون مجهولة ويرمز لها بالرمز π مساوية لقيمة محددة ويرمز لها بالرمز P_0 .

وهناك عدة حالات تتعلق بإجراء مثل هذا الاختبار ، إلا أن اختبارات النسبة في حالة العينات الصغيرة عادة ما تكون لها توزيعات احتمالية معقدة ، لذلك نفضل إجراء مثل هذه الاختبارات في وجود العينات الكبيرة وعندها فان التوزيع الاحتمالي لنسبة الحدوث في العينة سوف تقترب من التوزيع الطبيعي ، وبالتالي فان دالة الاختبار تأخذ الصورة التالية :

$$z = \frac{\hat{p} - p_0}{\text{SE}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \quad \text{where} \quad \hat{p} = \frac{x}{n}.$$

مثال :

أخذت عينة عشوائية مكونة من ٩٠٠ شخص فوجد أن عدد المؤيدين منهم لرأي معين ٧٣٨ شخص ، اختبر مدى صحة الفرضية القائلة بأن نسبة المؤيدين لذلك الرأي في المجتمع المسحوب منه العينة هي ٠٫٨ عند مستوى معنوية ٠٫٠٥ .

$n=900$, $P_0= 0.8$, $\alpha = 0.05$

ويمكن تقدير P حيث :

$$P = 738 / 900 = 0.82$$

$$Z = \frac{0.82 - 0.8}{\sqrt{\frac{0.8 * 0.2}{900}}} = 0.02 / 0.01333 = 1.5$$

وحيث أن قيمة دالة الاختبار Z المحسوبة من بيانات العينة تقع في منطقة رفض لفرض العدمي ، إذن نرفض الفرض العدمي ونقبل البديل القائل بأن نسبة المؤيدين لذلك الرأي هي 0.8 عندي مستوى معنوية 0.05 .

اختبار الفرضيات حول الفرق بين نسبتين في مجتمعين باستخدام بيانات عينتين مستقلتين :

في هذه الحالة أيضا يفضل استخدام عينات كبيرة لان توزيعات المعينة للنسبة تكون عادة معقدة في حالة العينات الصغيرة، لذلك نفضل إجراء مثل هذه الاختبارات في وجود العينات الكبيرة وعندها فان التوزيع الاحتمالي لنسبة الحدوث في العينة سوف تقترب من التوزيع الطبيعي ، وبالتالي فان دالة الاختبار تأخذ الصورة التالية :

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{SE}} = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \cong \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$$

$$= \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \quad \text{where} \quad \hat{p}_1 = \frac{x_1}{n_1},\ \hat{p}_2 = \frac{x_2}{n_2} \quad \text{and} \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2}.$$

مثال 2

أخذت عينة عشوائية من ١٢٢٢ شخص من منطقة أ ووجد أن عدد المتعلمين منهم ٣٦٤ شخص ، وأخذت عينة عشوائية ثانية مستقلة عن الأولى مكونة من ٣٢٧ شخص من منطقة ب ، ووجد ان عدد المتعلمين منهم ١٥٠ شخص ، اختبر ما إذا كان هناك فرق بين نسبتي المتعلمين في المنطقتين عند مستوى معنوية ٠,٠٥ .

الحل

For the data in **Example 2**, $\hat{p}_1 = 364/1222 = 0.30$, $\hat{p}_2 = 150/327 = 0.46$ and $\hat{p} = (364+150)/(1222+327) = 0.33$. Thus, $\text{SE} = \sqrt{0.33(0.67)\left(\frac{1}{1222} + \frac{1}{327}\right)} = 0.03$

$$z = (0.30 - 0.46)/0.03 = -5.33$$

وحيث أن قيمة دالة الاختبار Z المحسوبة من بيانات العينة تقع في منطقة رفض لفرض العدمي ، إذن نرفض الفرض العدمي ونقبل البديلة القائلة بأنه يوجد فرق جوهري بين نسبتي المتعلمين في المنطقتين ، عندي مستوى معنوية 0.05.

اختبار الفرضيات حول تباين مجتمع واحد

وفي هذه الحالة يكون الاهتمام منصبا على اختبار فرضيات حول التباين كمقاس تشتت لمجتمع واحد ، فمثل هذا الاختبار يساعد في تفسير التشتت في البيانات الإحصائية التي تكشف عن الفروق الفردية .
ولاختبار صحة الفرض العدمي نستخدم دالة الاختبار التالية :

$$X^2_{n-1} = \frac{(n-1)\, S^2}{\sigma^2_0}$$

مثال :
أخذت عينة عشوائية مكونة من ١٦ طالب وسجلت درجاتهم في أحد الامتحانات ، فوجد أن الانحراف المعياري المحسوب من العينة هو ٧ درجات فإذا كان σ^2 ترمز لتباين درجات الطلاب في هذه الامتحان فالمطلوب اختبار الفرضية العدمية التالية ، عند مستوى معنوية ٠,٠١ .

$H_0: \sigma^2 = 36$
$H_0: \sigma^2 > 36$

الحل
من البيانات المعطاة في هذا المثال نلاحظ أن :

$n=16 \qquad S^2=16 \qquad \sigma_0^2=36 \qquad \alpha = 0.01$

وان دالة الاختبار للفروض هي

$$\chi^2_{(n-1)} \qquad \frac{(n-1)s^2}{\sigma^2} \qquad = (15 * 49)/36 = 20.417$$

وهذه تتبع توزيع X^2 بدرجات حرية $16 - 1=15$.

وحيث أن X^2 المحسوبة من بيانات العينة 20.417 اصغر من القيمة الحرجة 30.578 أي تقع في منطقة القبول ، فانه لا يوجد دليل كافي لرفض الفرضية العدمية عند مستوى معنوية 0.01 .

اختبار X^2 لجودة التوفيق

يعطى الإحصاء X^2 التالي مقياسا لمدى التفاوت الوجود بين التكرارات المتوقعة في ظل صحة فرضية معينة والتكرارات المشاهدة من بيانات العينة ويعرف كالأتي :

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(O_{ij}-\hat{E}_{ij}\right)^2}{\hat{E}_{ij}}, \quad \text{where} \quad \hat{E}_{ij}=\frac{R_i C_j}{n},$$

مثال

ألقيت زهرة نرد ٦٠٠ مرة ، وظهرت الأرقام من ١ - ٦ بالتكرارات المبينة في الجدول :

الرقم	١	٢	٣	٤	٥	٦
التكرارات	١٠٠	٩٤	١٠٣	٨٩	١١٠	١٠٤

بمستوى معنوية ٠,٠١ أختبر الفرضية القائلة بأن زهرة النرد متوازنة .

لاختبار أن الزهرة متوازنة سوف نختبر الفرضية القائلة بأن القيمة التي تظهر على الزهرة تتبع التوزيع المنتظم باحتمال $1/6$ لكل قيمة ، وبالتالي يمكن صياغة الفرضية العدمية على النحو التالي :

$H_0 = P1 = P2 = ... = P6 = 1/6$

$H_a =$ ليست جميع الاحتمالات متساوية

وبالتالي في ظل صحة الفرضية العدمية سوف تظهر كل قيمة من القيم الستة بتكرار متوقع $n * P_j = 600 * 1/6 = 100$

$$X^2 = \frac{(100 - 100)^2 + (94 - 100)^2 + ... + (104 - 100)^2}{100} = 2.82$$

وحيث أن هناك ٦ خلايا في الجدول ولم يتم تقدير أي معلمة فان هناك ٥ درجات حرية .

ومن جدول توزيع كآي تربيع بخمس درجات حرية نجد أن قيمة كآي تربيع الجدولية بدرجات حرية ٥ تساوي 15.0863 وذلك عند مستوى معنوية 0.01 .

وبما أن دالة الاختبار المحسوبة كآي سكوير تقع في منطقة قبول الفض العدمي إذن نقبل الفرض العدمي القائل بان زهرة النرد متوازنة .

الفصل الرابع

مقارنة المتوسطات

Comparison of Means

- اختبارات t للمقارنة بين متوسطين
  - اختبار t للعينات المترابطة
  - أختار t لعينتين مستقلتين
  - اختبار t لعينة واحدة
- الإشارة للعينة الواحدة
- اختبار ويلكوكسون للرتب المؤشرة للعينة الواحدة
- توزيع F
- تحليل التباين في اتجاه واحد

أولاً : اختبار t للعينات المترابطة Paired-Samples T-test

وهنا يكون لدينا مجموعة واحدة تم قياس المتغير لديها مرتين ولذلك لكل فرد قيم متناظرة أو متزاوجة في مرتي القياس فمثلا تم تقدير درجات مجموعة من التلاميذ في العدوانية (المتوسط الأول) وبعد إخضاع العينة لبرنامج إرشادي تم قياس القلق مرة أخرى (المتوسط الثاني) فبالتالي يكون لكل فرد درجتين متناظرتين ، أو مثلا تم أخذ قراءة لتركيز عنصر ما في النبات قبل وبعد تعريضه للضوء فهنا نكون بصدد قراءات متناظرة لكل نبات أو تم قياس متغير معين لدى مجموعة من الذكور كبار السن قبل وبعد إعطائهم عقار معين المهم هذه الحالة تتعامل مع (عينة واحدة تم القياس عليها مرتين مختلفتين)
ولإجراء الاختبار نستخدم القانون التالي :

$$t = \frac{\overline{x}_d - c}{s_d / \sqrt{n}}.$$

مثال:
قام احد الباحثين بإنشاء برنامج الكتروني جديد لمساعدة الطلاب في دراسة مادة الأحياء وإجراء تجربه لمقارنة درجات الطلاب (مجموعه واحده قبل وبعد) لمعرفة هل هناك فرق بين درجات مجموعة الطلاب قبل استخدام البرنامج وبعد استخدام البرنامج وكانت درجات الطلاب كما يلي :

قبل	٦٠	٥٥	٦٥	٥٩	٤٢	٤٨	٦١	٣٧	٢٩	٦٢
بعد	٧٥	٦٨	٧٢	٦٠	٥٣	٥٩	٦٩	٤٥	٤٤	٧٤

وباستخدام برنامج التحليل الإحصائي SPSS نجد أن :

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	before - after	-10.100-	4.254	1.345	-13.143-	-7.057-	-7.507-	9	.000

وبما أن قيمة الـ p-value Sig اقل من 0.05 إذن نرفض الفرض العدمي ونقبل البديل القائل بوجود اختلاف بين العلامات .

ثانيا : أختار t لعينتين مستقلتين Independent-Samples T-test

وهـي أكثـر الحـالات اسـتخداماً والـتي فيهـا يـتم المقارنـة بـين متوسـطين مجمـوعتين مختلفـتين (الـذكور والإنـاث في الـذكاء مـثلا أو في الابتكـار أو في الـوزن أو في التحصيل) أو متوسطي الـدخل لشركتين أو قوة تحمـل الـضغوط لدى الذكور والإناث أو الرضا عن العمل لـدى مجموعتين مـن عمـال المصانع المهم من الضروري مراعاة وجود مجموعتين مختلفتين .

وتكون الفرضية العدمية والفرضية البديلة كالتالي :

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2.$$

ولإجراء الاختبار نستخدم القانون التالي :

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

حيث

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}.$$

ثالثاً: اختبار t لعينة واحدة One-Sample T-test

هـذه الحالـة تعـد مـن الحـالات الخاصـة جـداً لاختبـار " t " وفيهـا يـتم مقارنـة متوسط عينة ما (عينة واحدة) بمتوسط مجتمع معروف فمـثلا تم أخذ عينـة مـن إنتـاج مـصنع معـين وتم حساب متوسط الـوزن لهـا فكـانٍ ٥٣ ، مـع العلـم بـأن متوسط الوزن في المجتمع ٥٦ فهل هناك فروق دالة إحصائيا بين متوسط العينة والمتوسط المراد مقارنته به .

تفسير النواتج

في كل حالات T-test يتضمن ملف النتائج قيمة " t " ودلالتها الإحصائية Sigودرجات الحرية df والفروق بين المتوسطين Mean-Difference وهى أهم النواتج وإذا كانت القيمة في خانة .Sig أكبر من ٠,٠١ وأقل من ٠,٠٥ تكون هناك فروق بين المتوسطين في صالح المتوسط الأكبر أما إذا كانت القيمة أقل من ٠,٠١ وأكبر من ٠,٠٠١ تكون الفروق دالة عند ٠,٠١ أما إذا كانت القيمة أقل من أو تساوي ٠,٠٠١ تكون الفروق دالة عند مستوى ٠,٠٠١ .
ويجب ملاحظة أن df : في الحالة الأول (عينة واحدة) تساوي حجم العينة ناقص واحد وفي الحالة الثانية (مجموعتين) تساوي مجموع العينتين ناقص ٢ أو كل المجموع الكلي للحالات ناقص ٢ وفي الحالة الأخيرة (عينة واحدة) تساوي حجم العينة ناقص واحد.

وبالتالي في النهاية يكون القرار إما قبول الفرض الصفري Two means are not significantly different في حالة ما إذا كانت القيمة في خانة .Sig أكبر من ٠,٠٥ أو رفض الفرض الصفري Two means are significantly different في حالة القيمة في خانة .Sig أقل من أو تساوي ٠,٠٥.

اختبار الإشارة للعينة الواحدة : One sample sign test

يستخدم اختبار الإشارة للعينة الواحدة عادة للاستدلال على وسيط المجتمع خاصة عندما يكون توزيع المجتمع غير متماثل ، وفي حالة تماثل المجتمع فإن الاستدلال على الوسيط هو استدلال على المتوسط. ويشكل الاختبار في هذه الحالة بديلاً لامعلمياً لاختبار t للمتوسط والاختبار الطبيعي للمتوسط في الحالة التي يكون فيها التباين معروفاً. ورغم أنه أبسط – من حيث الحسابات المطلوبة – من اختبار ويلكوكسون ، إلا أنه أقل كفاءة منه. إذ اتضح أن الكفاءة النسبية لاختبار الإشارة بالنسبة لاختبار t (إذا كان توزيع المجتمع طبيعياً) هي فقط في حدود ٠,٦٣٧ بينما هي لاختبار ويلكوكسون ٠,٩٥ .
والبيانات التي يطبق عليها اختبار الإشارة يجب أن تكون مقاسه بالمقياس الترتيبي كحد أدنى حيث الترتيب بالنسبة للوسيط المفترض.

ولتوضيح الفكرة وراء هذا الاختبار نفرض أننا نريد أن نختبر ما إذا كان وسيط المجتمع يساوي المقدار mo وذلك باستخدام العينة العشوائية المستقلة X1, X2, ..., Xn المأخوذة من ذلك المجتمع. وحتى نتفادى أن تكون هناك مشاهده مساوية لـ mo تحدث ما يسمى بربطه سنضيف افتراضا آخر هو أن توزيع المجتمع متصل أو هو متصل في جوار m o . فإذا رمزنا لوسيط المجتمع بـ m فإن فرض العدم المطلوب اختباره هو

$$H_0 : m = m_0$$

و الفرض البديل هو

$$H_1 : m \neq m_0$$

اختبار ويلكوكسون للرتب المؤشرة للعينة الواحدة :

Wilcoxon (one sample) signed-rank test.

رأينا أن اختبار الإشارة يتميز بسهولة تطبيقه، كما أن اعتماده على الإشارات فقط قد يجعله الخيار الوحيد المتاح إذا كانت البيانات معطاة في شكل رتب أو أسماء أو إشارات تحدد رتبا بالنسبة للوسيط المفترض. غير أنه إذا كانت البيانات مقاسه بمقياس فوق الترتيبي (فترة أو نسبة) فإن استخدام اختبار الإشارة يؤدي لضياع معلومات متاحة لأنه لا يستفيد من مقدار الاختلاف بين القيم وإنما فقط موقع كل منها بالنسبة للوسيط. في مثل هذه الحالات نحتاج لاختبار يضع مقادير الاختلافات بين القيم في الاعتبار ويكون بالتالي أكثر كفاءة من اختبار الإشارة.

ولاختبار فرض العدم بأن وسيط المجتمع يساوي mo سنفترض أن لدينا العينة العشوائية المستقلة X1 , X2 , ... , Xn* المأخوذة من توزيع متصل (أو متصل في جوار mo) ومتماثل حول mo . أما البيانات فسنفترض أنها مقاسه بمقياس فترة كحد أدنى

مثال

في الجدول التالي يوضح العمود الثاني ، الثالث ، الرابع والخامس الفروقات ، والفروقات المطلقة، رتب الفروقات المطلقة والرتب المؤشرة بالترتيب. وعند إيجاد رتب $|di|$ بالعمود الرابع نلاحظ أن أصغر قيمة لـ $|di|$ هي ٠.٠١ وقد تكررت ثلاث مرات. ولو كانت هذه القيم مختلفة لكان ترتيبها ١ ، ٢ و٣ لهذا نعطى كل قيمة ٠,٠١ الرتبة

$$\frac{1+2+3}{3} = 2$$

تلي ٠,٠١ القيمة ٠,٠٢ وهي تكررت أربع مرات ولو كانت مختلفة لكان ترتيبها ٤، ٥، ٦،٧ لهذا نعطى كل ٠,٠٢ الرتبة

$$\frac{4+5+6+7}{4} = 5.5$$

وهكذا . وبالطبع إذا كانت القيمة غير متكررة لا تقاسمها الرتبة قيمة أخرى.

Xi	di = Xi - 2	|di|	رتبه |di|	الرتبة المؤشرة
١,٩٨	−٠,٠٢	٠,٠٢	٥,٥	−٥,٥
١,٩٨	−٠,٠٢	٠,٠٢	٥,٥	−٥,٥
٢,١٠	+٠,١٠	٠,١٠	١٣	+١٣
١,٩٩	−٠,٠١	٠,٠١	٢	−٢
١,٩٧	−٠,٠٣	٠,٠٣	٨,٥	−٨,٥
١,٩٨	−٠,٠٢	٠,٠٢	٥,٥	−٥,٥
٢,٠٠	٠	٠		
٢,١٢	+٠,١٢	٠,١٢	١٤	+١٤
٢,٠٩	+٠,٠٩	٠,٠٩	١٢	+١٢
١,٩٩	−٠,٠١	٠,٠١	٢	−٢
١,٩٨	−٠,٠٢	٠,٠٢	٥,٥	−٥,٥
١,٩٩	−٠,٠١	٠,٠١	٢	−٢
٢,٠٣	+٠,٠٣	٠,٠٣	٨,٥	+٨,٥
١,٩٦	−٠,٠٤	٠,٠٤	١٠	−١٠
١,٩٥	−٠,٠٥	٠,٠٥	١١	−١١

مثلاً القيمة ٠,٠٤ تأتي العاشرة في الترتيب وليست هناك قيمة أخرى تساويها لهذا أخذت الرتبة ١٠ .

بعد حذف القيمة التي تساوي الوسيط وهي ٢ يصبح حجم العينة $n = 14$.

مجموع الرتب الموجبة :

$$T+ = 13 + 14 + 12 + 8.5 = 47.5$$

مجموع الرتب السالبة :

$$T- = 5.5 + 5.5 + 2 + 8.5 + 5.5 + 2 + 5.5 + 2 + 10 + 11 = 57.5$$

لاحظ أن مجموع $T+$ و $T-$ يساوي ١٠٥ وهو مجموع الأعداد الطبيعية الـ ١٤ الأولى الذي يمكن أن نحصل عليه من القاعدة المعروفة لمجوع الأعداد الطبيعية:

$$\frac{14 \times 15}{2} = 105$$

هذا يمثل اختباراً لصحة الحسابات .

بما أن الاختبار ذو طرفين نأخذ أكبر المجموعين وهو ٥٧,٥ قيمة T- .
من جدول نجد عند $n = 14$ و $T = 57$ بالعمود المشار إليه بـ" يمين " أن $P^* = 0.404$ كما نجد عند $n = 14$ و $T = 58$ بنفس العمود $P^* = .380$.
وبأخذ متوسط ٠,٤٠٤ و ٠,٣٨٠ نحصل على تقدير لـ P^* الخاصة بـ ٥٧,٥ وهو ٠,٣٩٢ . وبالتالي فإن $P = 2 \times 0.392 = 0.784$. وبما أن هذا الاحتمال أكبر بكثير من ٠,٠٥ إذن نقبل الفرض العدمي .

توزيع F

ليكن لدينا المتغيران العشوائيان المستقلان X_1 ، X_2 وكل منهم يتبع توزيع معتدل وأخذنا عينة عشوائية حجمها n_1 من المجتمع X_1 ، عينة عشوائية حجمها n_2 من المجتمع X_2 ثم حصلنا على تقدير غير متميز لتباين المجتمع الأول σ_1^2 هوS_1^2 وكذلك للمجتمع الثاني ل σ_2^2 هو S_2^2 فالنسبة بين التباينين تتبع توزيع f بدرجات حرية لكل من البسط والمقام $V1=n1- 1$ (للبسط) ، $V2=n2- 1$ (للمقام) وهذا التوزيع يستخدم لاختبار تساوي مجتمعين وإحصائية الاختبار F هي: $F = S_1^2 \div S_2^2$ والتوزيع هذا يختلف باختلاف درجات الحرية سواء للبسط أو المقام أو كلاهما فمثلاً نرمز للتوزيع f بدرجة حرية للبسط 4 ، ودرجة حرية للمقام 8 عند مستوى معنوية 0.05 بالصورة:

$$F(\alpha , V1 , V2)$$

$$F(0.05 , 4 , 8)$$

يجب التأكيد على أنه إذا كانت النسبة أقل من الواحد الصحيح (معنوياً) فهذا يدل على تساوي تباين المجتمعين وإلا يوجد دلالة على اختلاف حقيقي بين تبايني المجتمعين أما قبول أو رفض فرض العدم ينتج من مقارنة قيمة F المحتسبة مع قيمة Fالجدولية.

تحليل التباين في اتجاه واحد One Way Analyses of Variance

هو طريقة لاختبار معنوية الفرق بين المتوسطات لعدة عينات بمقارنة واحدة، ويعرف أيضاً بطريقة تؤدي لتقسيم الاختلافات الكلية لمجموعة من المشاهدات التجريبية لعدة أجزاء للتعرف على مصدر الاختلاف بينها ولذا فالهدف هنا فحص تباين المجتمع لمعرفة مدى تساوى متوسطات المجتمع ولكن لا بد من تحقيق ثلاثة أمور قبل استخدامه وهي:

١) العينات عشوائية ومستقلة.

٢) مجتمعات هذه العينات كلاً لها توزيع طبيعي.

٣) تساوي تباين المجتمعات التي أخذت منها العينات العشوائية المستقلة.

ولتوضيح ما سبق بمقارنة متوسطات ثلاث مجتمعات باستخدام ثلاث عينات (تحقق فيها الشروط الثلاثة السابقة) موضحة بالجدول الآتي:

السؤال: هل في البيانات ما يكفي لوجود فرق بين المتوسطات؟

الجواب: نعم (بمجرد النظر) فالتشتت (التباين) ظاهر ٤٠، ٢٧، ٣٢,٥ (المتوسطات) بمقارنته بالتشتت بين العيناتٍ (وحداتها ٤٠ ، ٤١ ، ٤٠,٥، ٣٨,٥) فيبدو معدوما.

العينة الأولى	العينة الثانية	العينة الثالثة
40	27	33
41	28	32
40.5	26.5	33.5
38.5	26.5	31.5
$\bar{X}_1 = 40$	$\bar{X}_2 = 27$	$\bar{X}_3 = 32.5$
$S_1 = 1.08$	$S_2 = 0.71$	$S_3 = 0.91$

إذا أخذنا البيانات الآتية:

فالبيانات هنا لها نفس المتوسطات في البيانات السابقة ولكن التشتت (داخل لعينات) كبيراً بما هو عليه في المتوسطات.

العينة الأولى	العينة الثانية	العينة الثالثة
40	50	10
15	20	60
65	11	27.5
$\bar{X}_1 = 40$	$\bar{X}_2 = 27$	$\bar{X}_3 = 32.5$
$S_1 = 25$	$S_2 = 20.4$	$S_3 = 25.4$

فالدليل على وجود الفرق بين متوسطات الجدول الأول واضح ولا يظهر ذلك بوضوح في بيانات الجدول الثاني بالرغم من تساوي المتوسطات في الحالتين ولذا يتبين لنا القصد من تحليل التباين والذي يعني الفرق بين المتوسطات والذي يقاس بالتشتت داخل البيانات.

مثال:

في دراسة لتأثير وجود الطلاب في الصفوف على تحصيلهم في مادة الإحصاء، قام أستاذ الإحصاء بأخذ عينات عشوائية ومستقلة من ثلاثة صفوف (يقوم بتدريسها) كل منها مكون من خمسة طلاب وقام الأستاذ برصد درجاتهم والجدول التالي يبينها. بمستوى معنوية $\alpha = 0.05$ اختبر ما إذا كان متوسط النتائج في اختبارات الأداء يختلف في تحصيل الطلاب.

Class 1	Class 2	Class 3
66	96	58
65	87	62
88	66	77
92	55	90
60	78	80

الحل:

الاختبار:

$H_0: \mu_1 = \mu_2 = \mu_3$

H_a : على الأقل يوجد متوسطين غير متساويين

نستكمل الجدول كالآتي:

Class 1		Class 2		Class 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
66	4356	96	9216	58	3364
65	4225	87	7569	62	3844
88	7744	66	4356	77	5929
92	8464	55	3025	90	8100
60	3600	78	6084	80	6400
$T_1 = 371, T_1^2 = 137641$	28389	$T_2 = 382, T_2^2 = 145924$	30250	$T_3 = 367, T_3^2 = 134589$	27637

$T = T_1 + T_2 + T_3 = 371 + 382 + 367 = 1120 \quad , \quad T^2 = 1254400 \quad ,$

$n_1 = n_2 = n_3 = 5 \quad , \quad N = 15$

$SSB = 137641 / 5 + 145924 / 5 + 134689 / 5 – 1254400 / 15$

$= 418254 / 5 – 1254400 / 15$

$= 83650.8 – 83626.7$

$= 24.1$

$SSW = \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - 83650.5$

$= 28389 + 30250 + 27637 - 83650.5$

$= 86276 - 83650.5$

$= 2625.5$

$S_B^2 = 24.1 / (3 - 1) = 12.1$

$S_W^2 = 2625.5 / (15 - 3) = 218.8$

$F = S_B^2 / S_W^2$

$F = 12.05 / 218.8$

$F = 0.055 < 3.89 = F_{\alpha(2, 12)}$

مصدر التباين Source of Variance	مجموع المربعات Sum of squares (SS)	درجات الحرية df	متوسط مجموع المربعات أو التباين Mean squares (MS)	F (المحسوبة) Calculated	F (الجدولية) Tabulated
بين المجموعات Between Groups	$SS_B = 24.1$	$K - 1 = 3 - 1 = 2$	$S_B^2 = 24.1/2 = 12.05$	S_B^2 / S_W^2 12.05/218.5 0.055	$F_{\alpha (K-1), (N-K)}$ 3.89
داخل المجموعات Within Groups (Error)	$SS_W = 2625.5$	$N - K = 15 - 3 = 12$	$S_W^2 = 2625.5/12 = 218.8$		
المجموع Total	$SS_T = SS_B + SS_W = 2649.6$	$N - 1 = 15 - 1 = 14$			

إن قيمة F المحسوبة أقل من قيمة F الجدولية ،ولذا نقبل الفرضية الصفرية عند α = ٠,٠٥ بعدم وجود اختلاف بين المتوسطات.

الفصل الخامس

الارتباط والانحدار الخطي البسيط

Correlation and Simple Linear Regression

- معامل ارتباط بيرسون
- الانحدار الخطى البسيط
- نموذج الانحدار الخطي
- تقدير نموذج الانحدار الخطي البسيط

معامل ارتباط بيرسون

معامل بيرسون للارتباط الخطي من أكثر معاملات الارتباط استخدامًا خاصة في العلوم الإنسانية و الاجتماعية.
و مستوى القياس المطلوب عند تطبيق معامل بيرسون للارتباط هو أن يكون كلا المتغيرين مقياس فترة أو نسبي أو بمعنى أخر أن تكون بيانات كلا المتغيرين الظاهرتين بيانات كمية.

يمكن حساب معامل بيرسون بدلالة القراءات لبيانات المتغيرين x , y باستخدام الصيغة التالية:

$$\hat{\rho} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

مثال :

فيما يلي المساحة المنزرعة بالأعلاف الخضراء بالألف هكتار، وإجمالي إنتاج اللحوم بالألف طن، خلال الفترة من 1995حتى عام 2002.

السنة	1995	1996	1997	1998	1999	2000	2001	2002
المساحة	305	313	297	289	233	214	240	217
الكمية	592	603	662	607	635	699	719	747

والمطلوب: حساب معامل الارتباط بين المساحة والكمية، وما هو مدلوله ؟

الحـل

بفرض أن (x) هي المساحة المنزرعة، (y) هي الكمية

- حساب الوسط الحسابي لكل من المساحة، والكمية $(\bar{x}\ ,\ \bar{y})$.

$$\bar{x} = \frac{\sum x}{n} = \frac{2108}{8} = 263.5\ ,\ \bar{y} = \frac{\sum y}{n} = \frac{5264}{8} = 658$$

- حساب المجاميع

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
305	592	41.5	1722.25	-66	4356	-2739
313	603	49.5	2450.25	-55	3025	-2722.5
297	662	33.5	1122.25	4	16	134
289	607	25.5	650.25	-51	2601	-1300.5
233	635	-30.5	930.25	-23	529	701.5
214	699	-49.5	2450.25	41	1681	-2029.5
240	719	-23.5	552.25	61	3721	-1433.5
217	747	-46.5	2162.25	89	7921	-4138.5
2108	5264	0	12040	0	23850	-13528

$$\sum(x-\bar{x})^2 = 12040\ ,\ \sum(y-\bar{y})^2 = 23850\ ,$$
$$\sum(x-\bar{x})(y-\bar{y}) = -13528$$

إذا معامل الارتباط قيمته هي:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} = \frac{-13528}{\sqrt{12040}\sqrt{23850}}$$
$$= \frac{-13528}{(109.727)(154.434)} = \frac{-13528}{16945.619} = -0.798$$

إذن يوجد ارتباط عكسي قوي بين المساحة المنزرعة، وكمية إنتاج اللحوم.

الانحدار الخطي البسيط Simple Regression

إن الغرض من استخدام أسلوب تحليل الانحدار الخطي البسيط، هو دراسة وتحليل أثر متغير كمي على متغير كمي آخر، ومن الأمثلة على ذلك ما يلي:

- دراسة أثر كمية السماد على إنتاجية الدونم.
- دراسة أثر الإنتاج على التكلفة.
- دراسة أثر كمية البروتين التي يتناولها الأبقار على الزيادة في الوزن.
- أثر الدخل على الإنفاق الاستهلاكي.

وهكذا هناك أمثلة في كثير من النواحي الاقتصادية، والزراعية، والتجارية، والعلوم السلوكية، وغيرها من المجالات الأخرى.

نموذج الانحدار الخطي

في تحليل الانحدار البسيط، نجد أن الباحث يهتم بدراسة أثر أحد المتغيرين ويسمى بالمتغير المستقل أو المتنبأ منه، على المتغير الثاني ويسمى بالمتغير التابع أو المتنبأ به، ومن ثم يمكن عرض نموذج الانحدار الخطي في شكل معادلة خطية من الدرجة الأولى، تعكس المتغير التابع كدالة في المتغير المستقل كما يلي:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y : هو المتغير التابع (الذي يتأثر)

x : هو المتغير المستقل (الذي يؤثر)

β_0 : هو الجزء المقطوع من المحور الرأسي y ، وهو يعكس قيمة المتغير التابع في حالة انعدام قيمة المتغير المستقل x ، أي في حالة $x = 0$

β_1 : ميل الخط المستقيم $(\beta_0 + \beta_1 x)$ ، ويعكس مقدار التغير في y إذا تغيرت x بوحدة واحدة.

e : هو الخطأ العشوائي، والذي يعبر عن الفرق بين القيمة الفعلية y ، والقيمة المقدرة $\hat{y} = \beta_0 + \beta_1 x$ ، أي أن : $e = y - (\beta_0 + \beta_1 x)$ ، ويمكن توضيح هذا الخطأ على الشكل التالي لنقط الانتشار.

تقدير نموذج الانحدار الخطي البسيط

يمكن تقدير معاملات الانحدار $(\beta_1 \ , \ \beta_0)$ في النموذج باستخدام طريقة المربعات الصغرى، وهذا التقدير هو الذي يجعل مجموع مربعات الأخطاء العشوائية $\sum e^2 = \sum \left(y-(\beta_0+\beta_1 x)\right)^2$ أقل ما يمكن، ويحسب هذا التقدير بالمعادلة التالية:

$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} \ ,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

حيث أن $\bar{x}$ هو الوسط الحسابي لقيم x، $\bar{y}$ هو الوسط الحسابي لقيم y، وتكون القيمة المقدرة للمتغير التابع هو: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$، ويطلق على هذا التقدير " تقدير معادلة انحدار y على x.

مثال

فيما يلي بيانات عن كمية البروتين اليومي بالجرام التي يحتاجها العجل الرضيع، ومقدار الزيادة في وزن العجل بالكجم، وذلك لعينة من العجول الرضيعة حجمها **10**.

كمية البروتين	10	11	14	15	20	25	46	50	59	70
الزيادة في الوزن	10	10	12	12	13	13	19	15	16	20

المطلوب :

١- ارسم شكل الانتشار، وما هو توقعاتك لشكل العلاقة ؟

٢- قدر معادلة انحدار الوزن على كمية البروتين.

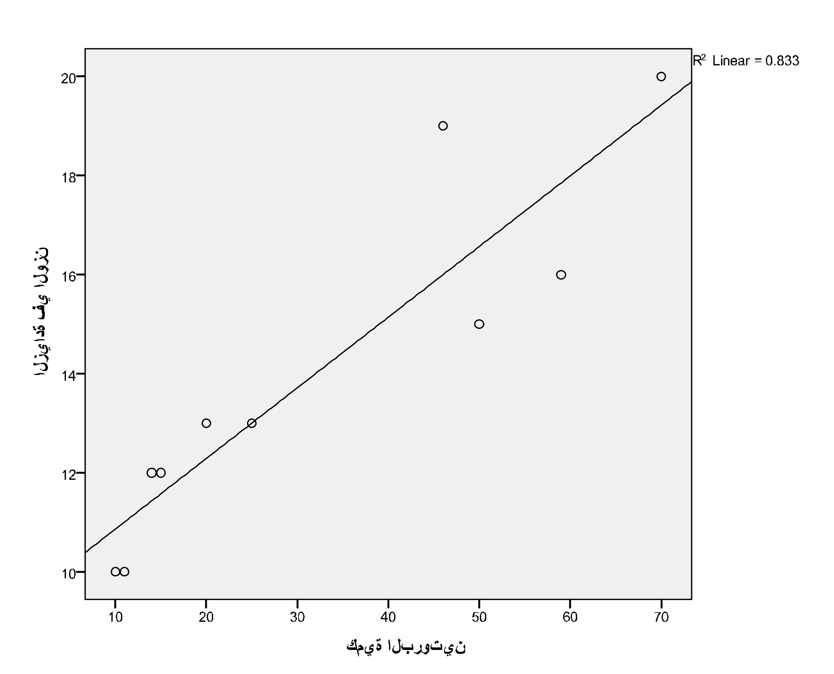
٣- فسر معادلة الانحدار.

٤- ما هو مقدار الزيادة في الوزن عند إعطاء العجل **50** جرام من البروتين ؟ وما هو مقدار الخطأ العشوائي؟

الحـل

١- رسم نقط الانتشار:





كمية البروتين X

من المتوقع أن يكون لكمية البروتين أثر طردي (إيجابي) على مقدار الزيادة في الوزن.

٢- تقدير معادلة الانحدار.

بفرض أن x هي كمية البروتين، y هي مقدار الزيادة في الوزن يتم حساب المجاميع التالية:

كمية البروتين x	الزيادة في الوزن y	$x\ y$	x^2
10	10	100	100
11	10	110	121
14	12	168	196
15	12	180	225
20	13	260	400
25	13	325	625
46	19	874	2116
50	15	750	2500
59	16	944	3481
70	20	1400	4900
320	140	5111	14664

المجاميع المطلوبة

$$\sum x = 320$$
$$\sum y = 140$$
$$\sum xy = 5111$$
$$\sum x^2 = 14664$$

إذا الوسط الحسابي:

$$\bar{x} = \frac{\sum x}{n} = \frac{320}{10} = 32$$
$$\bar{y} = \frac{\sum x}{n} = \frac{140}{10} = 14$$

$$\hat{\beta}_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{(10)(5111) - (320)(140)}{(10)(14664) - (320)^2}$$

$$= \frac{6310}{44240} = 0.1426$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 14 - (0.1426)(32) = 9.4368$$

إذا معادلة الانحدار المقدرة، هي:

$$\hat{y} = 9.44 + 0.143x$$

٣. تفسير المعادلة:

الثابت $\hat{\beta}_0 = 9.44$ **: يدل على أنه في حالة عدم استخدام البروتين في التغذية، فإن الوزن يزيد 9.44 كجم.**
معامل الانحدار $\hat{\beta}_1 = 0.143$ **: يدل على أنه كلما زادت كمية البروتين جرام واحد، حدث زيادة في وزن العجل بمقدار 0.143 كجم، أي زيادة مقدارها 143 جرام.**

٤. مقدار الزيادة في الوزن عند $x = 50$ **هو:**

$$\hat{y} = 9.44 + 0.143(50) = 16.59$$

وأما ومقدار الخطأ العشوائي هو:

$$\hat{e}_{x=50} = y_{x=50} - \hat{y}_{x=50} = 15 - 16.59 = -1.59$$

الفصل السادس

الانحدار الخطي المتعدد

Multiple Linear Regression

- الانحدار الخطي المتعدد
- طريقة المربعات الصغرى
- مثال تطبيقي باستخدام الـ SPSS

يهتم تحليل الانحدار الخطي المتعدد بدراسة وتحليل أثر عدة متغيرات مستقلة كمية على متغير تابع كمي .
نموذج الانحدار المتعدد هو عبارة عن انحدار للمتغير التابع (Y) على العديد من المتغيرات المستقلة $X_1 , X_2 , \ldots , X_K$ ويسمى هذا بنموذج الانحدار الخطي المتعدد Multiple Linear Regression .

يستند النموذج الخطي المتعدد على افتراض وجود علاقة خطية بين متغير تابع Y_i وعدد من المتغيرات المستقلة $X_1 , X_2 , \ldots , X_K$ وحد عشوائي U_i, ويعبر عن هذه العلاقة ، بالنسبة لـ n من المشاهدات و k من المتغيرات المستقلة ، بالشكل آلاتي :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$$Y_1 = B_0 + B_0 X_{11} + B_2 X_{12} + \ldots + B_K X_{1K} + U_1$$
$$Y_2 = B_0 + B_1 X_{21} + B_2 X_{22} + \ldots + B_K X_{2K} + U_2$$
$$\vdots$$
$$Y_n = B_0 + B_1 X_{n1} + B_2 X_{n2} + \ldots + B_K X_{nK} + U_n$$

هذه المعادلة تتضمن (k +١) من المعلومات المطلوب تقديرها علما بان الحد الأول منها (B_0) يمثل الحد الثابت ، الآمر الذي يتطلب اللجوء إلى المصفوفات والمتجهات لتقدير تلك المعلمات. عليه يمكن صياغة هذه المعادلات في صورة مصفوفات وكآلاتي :

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} \ldots x_{1k} \\ 1 & x_{21} & x_{22} \ldots x_{2k} \\ \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} \ldots x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = XB + E$$

حيث أن :

Y : متجه عمودي أبعاده (n +١) يحتوي مشاهدات المتغير التابع .

X: مصفوفة أبعادها $(n \times (k+1))$ تحتوي مشاهدات المتغيرات المستقلة يحتوي عمودها الأول على قيم الواحد الصحيح ليمثل الحد الثابت .

B: متجه عمودي أبعاده $((k+1)\times 1)$ يحتوي على المعالم المطلوب تقديرها .

U: متجه عمودي أبعاده (n ×١) يحتوي على الأخطاء العشوائية .

مثال :

بفرض لدينا المعطيات الآتية :

$N=16$, $\Sigma X_1 = 116$, $\Sigma X_2 = 48$, $\Sigma X_1^2 = 928$
ΣX_2^2 , $\Sigma X_1X_2 = 352$, $\Sigma Y = 1308$, $\Sigma X_1Y = 9862$
$\Sigma X_2Y = 3994$.

والمطلوب تقدير معادلة خط الانحدار باستخدام طريقة المربعات الصغرى ؟

$$\hat{B} = (X'X)^{-1}X'Y$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{bmatrix} n & \sum x_1 & \sum x_2 \\ \sum x_1 & \sum x_1^2 & \sum x_1x_2 \\ \sum x_2 & \sum x_1x_2 & \sum x_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y \\ \sum x_1y \\ \sum x_2y \end{bmatrix}$$

$$(X'X) = \begin{bmatrix} 16 & 116 & 48 \\ 116 & 928 & 352 \\ 48 & 352 & 160 \end{bmatrix} \quad X'Y = \begin{bmatrix} 1308 \\ 9862 \\ 3994 \end{bmatrix}$$

$$\hat{B} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 41.302 \\ 4.203 \\ 3.324 \end{pmatrix}$$

إذن معادلة الانحدار الخطي المتعدد هي :

$$\hat{y}_i = 41.302 + 4.203x_{i1} + 3.324x_{i2}$$

مثال تطبيقي باستخدام الـ SPSS

البيانات التالية تمثل العلاقة بين الكمية من سلعة معينة (Y) والعوامل المؤثرة علها وهي السعر (X1) ، دخل المستهلك (X2) بالدولار ، سعر السلعة البديلة (X3) .
وحسب النظرية الاقتصادية هناك علاقة بين المتغير المعتمد وهو الكمية المطلوبة والمتغيرات التفسيرية (المستقلة)الأخرى وهي (السعر ، الدخل ، سعر السلعة البديلة)

سعر السلعة البديلة X3	الدخل X2	السعر X1	الكمية Y	السنوات
10	400	9	40	1981
14	500	8	45	1982
12	600	9	50	1983
13	700	8	55	1984
11	800	7	60	1985
15	900	6	70	1986
16	1000	6	65	1987
17	1100	8	65	1988
22	1200	5	75	1989
19	1300	5	75	1990
20	1400	5	80	1991
23	1500	3	100	1992
18	1600	4	90	1993
24	1700	3	95	1994
21	1800	4	85	1995

يمكن معرفة الأثر أو العلاقة بين المتغيرات التفسيرية والمتغير المعتمد من خلال تقدير هذه العلاقة وبالشكل (شكل العلاقة) الأتي:

$$Y = f (X1, X2 , X3)$$

والمعادلة التقديرية حسب نموذج الانحدار الخطي المتعدد تكون وفق الأتي :

$$Y = a + bX1 + cX2 + dX3 + u$$

حيث تمثل :

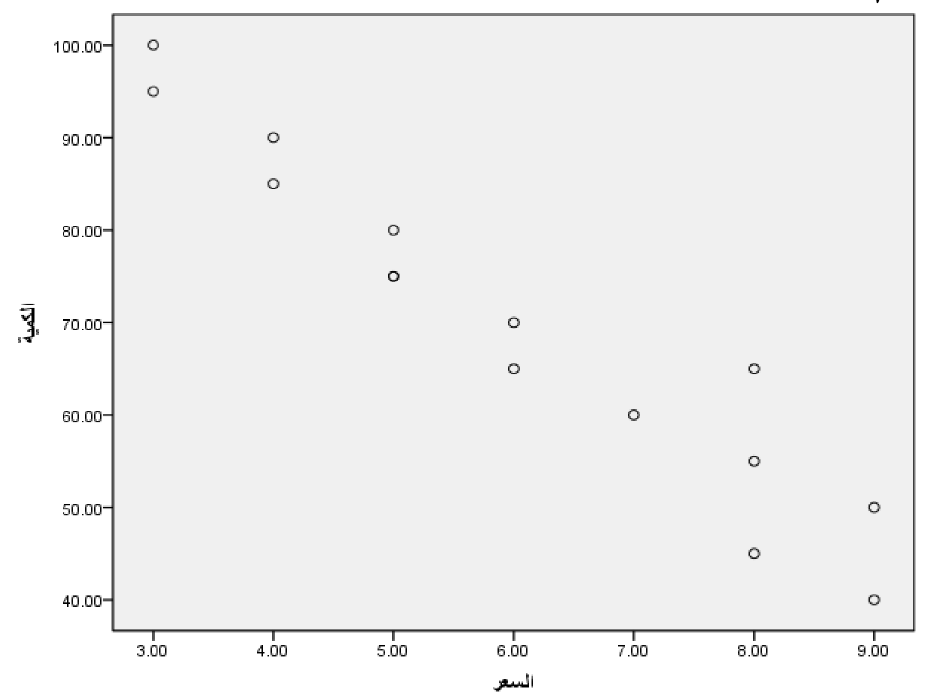
a : معامل التقاطع أو الحد الثابت .

b ، c ، d : تمثل معاملات معادلة الانحدار الخطي المتعدد .

U ، تمثل الخطأ القياسي أو الخطأ العشوائي للنموذج المقدر .

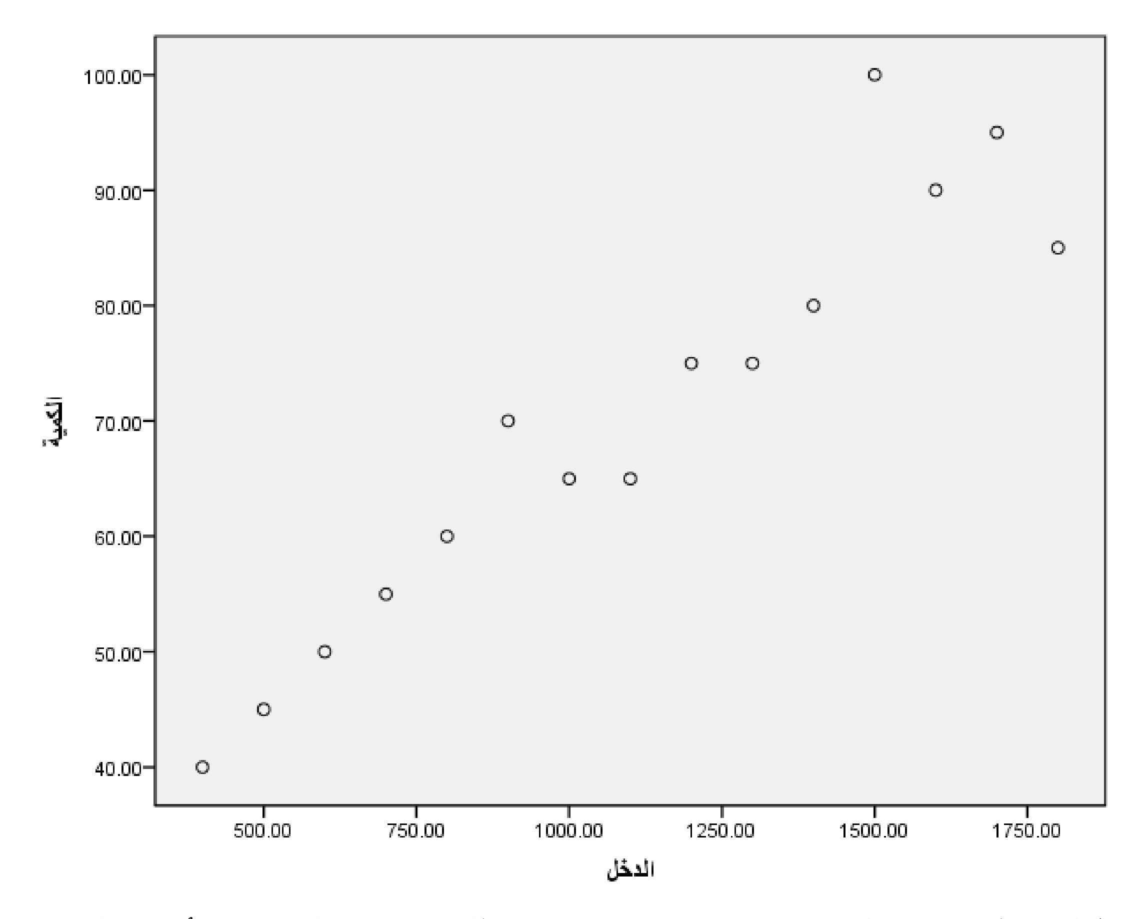
في البداية يمكن رسم شكل الانتشار بين كل من المتغير التابع (الكمية) وكل من المتغيرات المستقلة على حدا كما يلي :

أولا رسم شكل الانتشار بين الكمية المطلوبة والسعر :



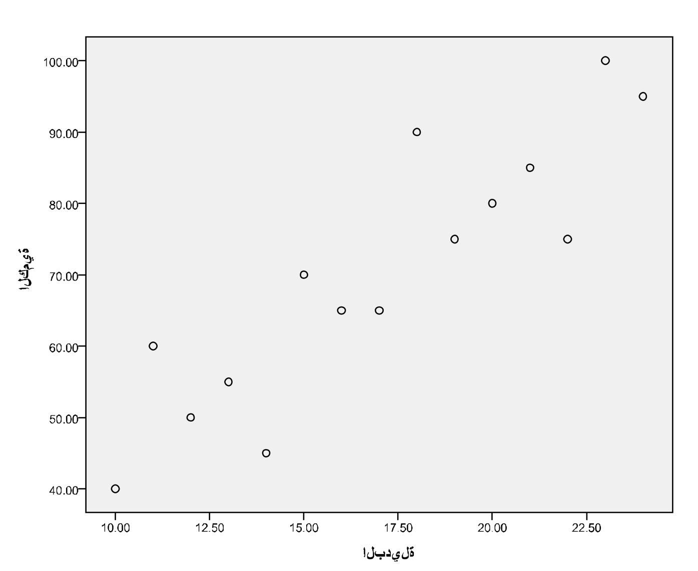
فيلاحظ من شكل الانتشار السابق بوجود علاقة عكسية بين السعر والكمية المطلوبة وهذا طبيعي .

أما برسم شكل الانتشار بين الكمية المطلوبة والدخل :



فيلاحظ وجود علاقة عكسية بين الكمية المطلوبة والدخل ، حيث أنه كلما زاد الدخل تزيد الكمية المطلوبة من السلع والخدمات وهذا طبيعي .

وبرسم شكل الانتشار بين كل من الكمية المطلوبة وأسعار السلع البديلة



حيث يلاحظ من الشكل بعدم وجود علاقة واضحة بين الكمية المطلوبة وأسعار السلع البديلة ، حيث أنه كلما زادت أسعار السلع البديلة انخفضت الكمية المطلوبة من هذه السلعة والعكس فانه كلما انخفضت أسعار السلع البديلة كلما زادت الكمية المطلوبة من السلع العادية .

وبطلب أمر الانحدار الخطي ومن ثم إدخال كل من المتغير التابع والمتغيرات المستقلة ، تظهر لدينا عدة جداول كما يلي :

<u>الجدول الأول</u> يمثل طريقة الانحدار المستخدمة وهي طريقة Enterحيث يتبين أن البرنامج قام بإدخال جميع المتغيرات المستقلة في معادلة الانحدار الخطي المتعدد.

<u>الجدول الثاني</u> : يوضح الجدول الثاني قيم معامل الارتباط الثلاثة وهي معامل الارتباط البسيط R حيث بلغ ٠,٩٧ ومعامل التحديد R^2 وهو يساوي ٠,٩٥ وأخيرا معامل التحديد المصحح R^2- والذي بلغ ٠,٩٤ مما يعني أن المتغيرات المستقلة (التفسيرية) (السعر ، الدخل ، سعر السلعة الأخرى) استطاعت ان تفسر ٠,٩٤ من التغيرات الحاصلة في الكمية المطلوبة (Y) والباقي (٠,٠٦) يعزى إلى عوامل أخرى.

<u>الجدول الثالث</u> : يمثل الجدول الثاني جدول تحليل التباين والذي يمكن المعرفة من خلاله على القوة التفسيرية للنموذج ككل عن طريق إحصائية F وكما نشاهد من جدول تحليل التباين المعنوية العالية لاختبار F (P < ٠,٠٠١) . مما يؤكد القوة التفسيرية العالية لنموذج الانحدار الخطي المتعدد من الناحية الإحصائية .

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4374.508	3	1458.169	71.133	.000[a]
	Residual	225.492	11	20.499		
	Total	4600.000	14			

a. Predictors: (Constant), البديلة, الدخل, السعر

b. Dependent Variable: الكمية

الجدول الرابع : يبين الجدول الرابع والأخير قيم معاملات الانحدار للمقدرات والاختبارات المعنوية الإحصائية لهذه المعاملات ويمكن تلخيص هذه الجدول بالشكل الأتي

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	79.106	19.782		3.999	.002
	السعر	-4.928-	1.611	-.563-	-3.059-	.011
	الدخل	.016	.007	.392	2.146	.055
	البديلة	.175	.637	.043	.275	.789

a. Dependent Variable: الكمية

من الجدول نستنتج أن المتغيرات المستقلة (سعر السلعة) كان معنوي من الناحية الإحصائية وحسب اختبار t عند مستوى معنوية $P \leq 0.05$ ، في حين كاد متغير الدخل ان يكون معنوي عند مستوى معنوية $P \leq 0.05.$ إلا أن المتغير المستقل سعر السلعة البديلة لم يكن ذو تأثير معنوي في نموذج الانحدار المتعدد وحسب اختبار t .

التفسير الاقتصادي :

حسب منطق النظرية الاقتصادية ، الكمية المطلوبة من سلعة معينة ترتبط بعلاقة عكسية مع السعر ، وبعلاقة طردية مع الدخل ، وبعلاقة طردية مع سعر السلعة البديلة ، ومن النتائج التي حصلنا عليها نجد أن جميع الإشارات كانت مطابقة مع النظرية الاقتصادية .

أن معامل السعر كان (٤,٩٣ -) وهذا مطابق لمنطق النظرية الاقتصادية ، مما يعني أن كل زيادة في السعر بمقدار دولار واحد سيؤدي إلى انخفاض الكمية المطلوبة بمقدار ٥ وحدات تقريبا (٤,٩٣) ، أما فيما يخص الدخل ، أيضا كان مطابق للنظرية الاقتصادية حيث كان (٦.٠١) مما يعني انه كل زيادة في الدخل بمقدار دولار واحد ستؤدي إلى ارتفاع الكمية المطلوبة بمقدار (١,٦) وحدة ، وأخيرا بالنسبة لعامل السلعة البديلة ، نجد ان ه أيضا مطابق للنظرية الاقتصادية حيث بلغت قيمته (٠,١٧) ، أي انه اذا ازداد الكمية المطلوبة من السلعة بمقدار وحدة واحدة فان الطلب على السلعة البديلة سوف يزداد بمقدار ٠,١٧ وحدة .؛

الفصل السابع

تصميم وتحليل التجارب

Design and Analysis of Experiments

- ❖ مفاهيم أساسية
- ❖ تصميم التجارب
- ❖ القواعد الأساسية في تصميم التجارب
- ❖ تصميم كامل العشوائية

تصميم وتحليل التجارب : هي فرع من فروع علم الإحصاء الذي يهتم بتطبيق الطريقة الإحصائية في التجربة العلمية .

سوف نستعرض في هذا الفصل عدة تصميمات مهمة في علم تصميم وتحليل التجارب ، منها :

١. تصميم كامل العشوائية .
٢. تصميم البلوكات أو القطاعات العشوائية .
٣. تصميم المربعات اللاتينية .

أولا :مفاهيم أساسية في تصميم التجارب basic concept in experimental design

التجربة : هو الطريقة العليمة التي تستخدم لاختبار الفرضيات واستكشاف العلاقات والمفاهيم التي تتعلق بمشكلة معينة من المشاكل . وبشكل آخر إن التجربة عمل منظم أو طريقة منتظمة لاستكشاف الحقائق والبراهين والفرضيات التي تتعلق بمشكلة معينة

الوحدات التجريبية : Experimental unit هي اصغر وحدة في التجربة ، وقد تكون الوحدات التجريبية أشخاص أو نباتات أو أراضي أو جزء من جسم الإنسان .

المعاملة : treatment مجموعة من الظروف التجريبية المتغيرة التي توضع تحت سيطرة الباحث والتي يقوم الباحث بتوزيعها على الوحدات التجريبية وفقا لنموذج المستخدم.

العامل: factor ويكون عبارة عن عنصر من عناصر المعاملة .

الخطأ التجريبي : Experimental Error هو مقياس للاختلافات الطبيعية التي توجد عادة بين مشاهدات سجلت من نفس الوحدات التجريبية التي عوملت بنفس المعاملة ، ولكي لا نقع في أشكال ، يتم اخذ الوحدات التجريبية المتجانسة (من نفس العمر والجنس والسلالة).

مصادر الخطأ التجريبي وتأتي من الأتي:

أ-**الاختلافات الذاتية** ، ترجع بالدرجة الأساس إلى الاختلافات الوراثية أو التداخل بين الاختلافات الوراثية والظروف البيئية التي يصعب السيطرة عليها
ب-**اختلافات في تطبيق المعاملة** .
ج-**الأخطاء الفنية** ، وتأتي من أخطاء عمليات القياس والتقدير.

ثانيا-تصميم التجارب Experimental design :

التصميم هو تخطيط التجربة مما يسهل جمع البيانات والمعلومات . ويجب أن يوضح التصميم قبل القيام بالتجربة .

وهنالك جملة اعتبارات يجب الأخذ بها عند التصميم :

١- عدد المشاهدات observations
٢- عدد العوامل factors
٣- هل العوامل ثابتة أم متغيرة

ما هو النموذج الرياضي المستخدم mathematical model

وعند وضع التصميم توضع ثلاثة أسئلة وهي :

أ- هل التصميم المطلوب هو من اجل تجربة من اجل عامل واحد uni factor أم ذات عوامل متعددة multi factors

ب-هل الوحدات التجريبية متجانسة .

ج-هل جميع المعاملات تتمثل في قطاعات متكاملة أم لا وهل أن تأثير العوامل تجميعي أم غير تجميعي وهل هو ثابت أم متغير.

ثالثا-القواعد الأساسية في تصميم التجارب

الأولى: العشوائية Randomization

ويقصد بها توزيع المعاملات على الوحدات التجريبية دونما تحيز . unbiased
وهذه القاعدة مهمة لعدة اعتبارات وهي:

تجنب الخطأ المنتظم
ضمان دقة تقدير الخطأ التجريبي وبالتالي فان كفاءة التجربة ونتائجها تكون صحيحة

الثانية : التكرار Replication

وهو من القواعد المهمة التي لا يمكن معرفة الخطأ التجريبي دونها وذلك لان التكرار هو مقياس للأخطاء التجريبية . ويحقق التكرار عدة فوائد هي

١- إمكانية تقدير قيمة الخطأ التجريبي
٢- زيادة كفاءة التجربة ودقتها لتقليل قيمة الخطأ التجريبي
٣- زيادة إمكانية تعميم ناتج التجربة .

الثالثة : التعرف على الوحدات التجريبية والتحكم بها .

رابعا : متطلبات التجربة الجيدة

١- غياب الخطأ المنتظم ويمكن تحقيق هذا الهدف من خلال استخدام مبدأ التوزيع العشوائي
٢- تقليل التباين بين الوحدات التجريبية . حيث كلما ازدادت عدد الوحدات التجريبية قل الخطأ التجريبي.
٣- استخدام التصميم المناسب للتجربة
٤- البساطة ، فكلما كان التصميم بسيط تكون التجربة أكثر جودة .

تصميم القطاعات العشوائية الكاملة

Complete Randomized Blocks Design

* الغرض من التصميم

يستخدم هذا التصميم إذا كانت الوحدات التجريبية غير متجانسة، بحيث أنه يمكن تقسيمها إلى قطاعات غير متجانسة، في حين أن كل قطاع يشمل مجموعة من الوحدات التجريبية المتجانسة تماما. ومن ثم يمكن تقليل تباين الخطأ التجريبي، كما سنوضح فيما بعد.

* التعشية

لبيان كيف تتم التعشية في هذا النوع من التصميم، يتم أخذ المثال التالي:

بفرض أن:

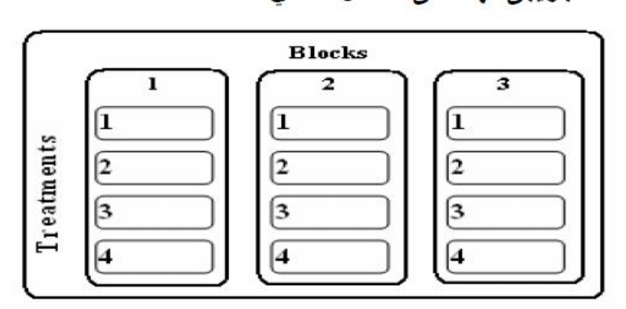
١- عدد المعالجات هو: $t = 4$ هي: (T_1,T_2,T_3,T_4)

٢- عدد القطاعات هو $b = 3$ هي: (B_1,B_2,B_3)

٣- أن كل خلية بها تكرار واحد

٤- فإن عدد الوحدات التجريبية هي: $t_. = b. = 3\times 4 = 12$ وحدة تجريبية.

٥- يكون المخطط التجريبى لها على الشكل التالي:

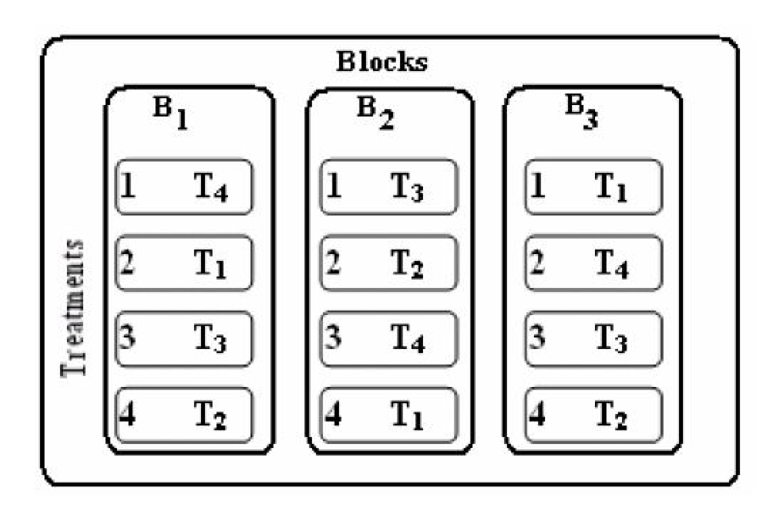


٦- وتتم التعشية للقطاعات، والمعالجات كالتالي:

- اختيار ثلاث أرقام عشوائية للقطاعات، ثم اختيار عدد 12 رقم عشوائي للمعالجات كما هو مبين بالجدول التالي:

Exper. Unit	R. Number for Treat.	Treatments	R. Number for Blocks	Blocks
1	09	T_1	53	B_3
2	91	T_4		
3	61	T_3		
4	39	T_2		
1	33	T_3	26	B_2
2	29	T_2		
3	97	T_4		
4	25	T_1		
1	82	T_4	14	B_1
2	32	T_1		
3	73	T_3		
4	39	T_2		

ومن ثم المخطط التجريبي هو:



*** النموذج الرياضي:**

يأخذ النموذج الرياضي في هذا التصميم شكل نموذج تحليل التباين الثنائي، مع تغيير المسميات:

بفرض أن القياسات تم تنظيمها على الشكل التالي:

		Treatments			
		1	2		t
Blocks	1	y_{11}	y_{12}		y_{1t}
	2	y_{21}	y_{22}		y_{2t}
	b	y_{b1}	y_{b2}		y_{bt}

حيث أن y_{ij} هي المشاهدة التي تقع تحت تأثير القطاع i، والمعالجة j ، ومن ثم يعبر عنها كما يلي:

$$y_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij} \ , \ i = 1,2,...,b \ , \ j = 1,2,...,t \qquad (1)$$

حيث أن:

y_{ij} **: هي المشاهدة التي تقع تحت تأثير القطاع i، والمعالجة j .**

μ **: هو المتوسط العام.**

β_i **: أثر القطاع رقم i ، $i = 1,2,...,b$ ، وهو يساوي:** $\beta_i = \mu_{i.} - \mu$.

τ_j **: أثر المعالجة رقم j ، $j = 1,2,...,t$ ، وهو يساوي:** $\tau_j = \mu_{.j} - \mu$

ε_{ij} **: هو الخطأ العشوائي للمشاهدة التي تقع تحت تأثير القطاع i، والمعالجة j .**

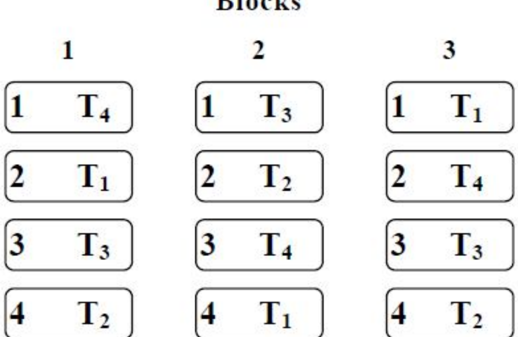
*** افتراضات النموذج**

$$\sum_{i=1}^{b} \beta_i = \sum_{j=1}^{t} \tau_j = 0$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

* جدول تحليل التباين:

S.O.V	d.f	SS	MS	F^*
Blocks	(b-1)	$SSB = \frac{1}{t}\sum_{i=1}^{b} Y_{i.}^2 - C.F$	MSB	$\frac{MSB}{MSE}$
Treatments	(t-1)	$SST = \frac{1}{b}\sum_{i=1}^{b} Y_{.j}^2 - C.F$	MST	$\frac{MST}{MSE}$
Errors	(b-1)(t-1)	$SSE = SSTo - (SSB + SST)$	MSE	
Total	tb-1	$SSTo = \sum_{i=1}^{b}\sum_{j=1}^{t} y_{ij}^2 - C.F$		



الفصل الثامن

الانحدار اللوجستي

Logistic Regression

- مفهوم الانحدار اللوجستي
- تحويلات الانحدار اللوجستي
- الاحتمال
- معامل الترجيح Odds
- تحويل معامل الترجيح Odds إلي دالة اللوجت Logit
- تفسير معاملات الانحدار اللوجستي
  1. تفسير المعاملات بدلالة اللوجت
  2. تفسير المعاملات بدلالة معاملات الترجيح

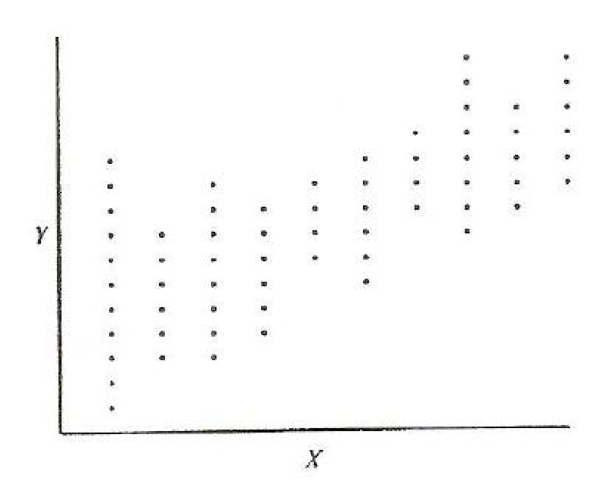
مفهوم الانحدار اللوجستي Logistic Regression

يعرّف الانحدار بشكل عام بأنّه التحليل الذي يختص بدراسة اعتماد متغير واحد يعرف بالمتغير التابع على متغير واحد أو أكثر يعرف بالمتغيّر المستقل أو المتغيرات المستقلة (المفسّرة) وذلك بغرض التقدير أو التنبؤ بمتوسط قيمة المتغيّر التابع بمعلومية المتغيرات المفسرة. وبناء على ذلك فإنّ أسلوب الانحدار يستخدم للتوصل إلى نموذج رياضي يوضح العلاقة الكمية بين المتغير التابع المراد التنبؤ بقيمته والمتغيرات المفسّرة

المشكلة المفاهيمية في استخدام انحدار المربعات الدنيا لتوفيق البيانات ذات المتغيرات التابعة الثنائية تنشأ من حقيقة أن الاحتمالات يجب أن تتراوح قيمها بين قيمتين حديتين هما: الواحد الصحيح كحد أعلى والصفر كحد أدنى، أي أنّه وفقاً لتعريف الاحتمالات لا يمكن لقيمة الاحتمال أن يتجاوز الواحد الصحيح، ولا أن ينخفض إلى ما دون الصفر. وحيث إنّ تحليل انحدار المربعات الدنيا هو نموذج خطي يسمح لخط الانحدار أن يمتد حتى موجب ما لا نهاية، أو أن يمتد حتى سالب ما لا نهاية حسب قيمة المتغيّر أو المتغيرات المستقلة، فإنّ استخدام انحدار المربعات الدنيا مع البيانات ذات المتغير التابع الثنائي قد يفاجئ الباحث بقيم متوقعة للمتغير التابع تتجاوز الواحد الصحيح أو تقل عن الصفر، الأمر الذي يتناقض تماماً مع مفهوم الاحتمالات.

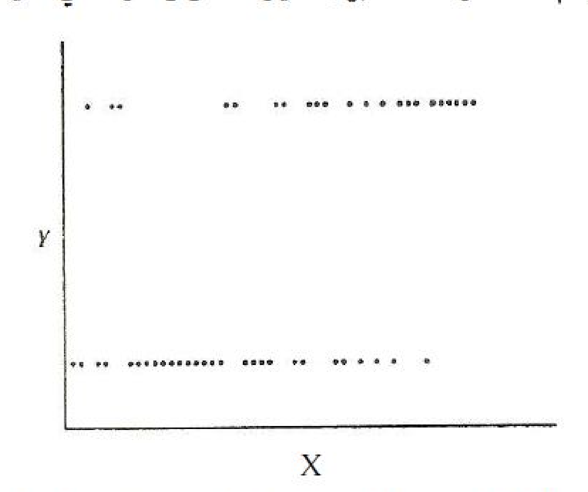
ولتوضيح الفكرة السابقة بيانياً يلاحظ أنّه عند تمثيل رسم الانتشار لمتغيرين متصلين، فإن رسم الانتشار سيكون على شكل نقاط تشبه السحابة، حيث يعتمد شكل تلك السحابة على قوة العلاقة بين المتغيرين المتصلين. وللتنبؤ بقيمة أحد المتغيرين وليكن Y بدلالة قيمة المتغير الآخر وليكن X، يتم رسم خط يمثّل أفضل توفيق للبيانات المشاهدة، بحيث يكون هذا الخط هو الذي يعبّر عن العلاقة بين المتغيرين المتصلين، بحيث يحقق خاصية أن مجموع مربعات انحرافات القيم المتوقعة (الواقعة على الخط) والقيم المشاهدة تكون أقل ما يمكن. وتسمى هذه الطريقة بالمربعات الدنيا. ويلاحظ في هذه الحالة ونظرياً على الأقل أن عملية التنبؤ بقيمة Y تتم باستخدام نفس الخط المستقيم، وأن ذلك الخط هو المعتمد عند التنبؤ بقيم Y بدلالة قيم المتغير X، سواء كانت قيمة X مرتفعة جداً، أو متوسطة، أو منخفضة.

شكل (1): رسم الانتشار للعلاقة بين متغيّرين متصلين



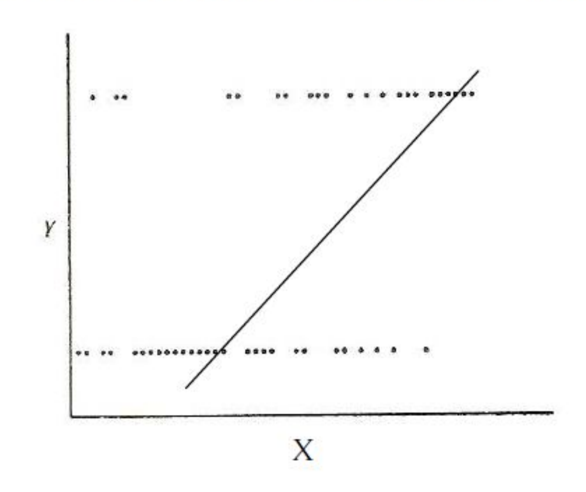
لكنّ الوضع يختلف قليلاً في حالة المتغير التابع الثنائي، حيث يلاحظ أن رسم الانتشار لا يظهر ما يشبه السحابة عند تمثيل العلاقة بين المتغير المستقل المتصل X والمتغير التابع الثنائي Y، بل إنّ رسم الانتشار في هذه الحالة هو عبارة عن مجموعتين من النقاط المتوازية.

شكل (2): رسم الانتشار للعلاقة بين متغيّر متصل وآخر ثنائي القيمة



وبذلك فإن محاولة رسم أفضل خط مستقيم لتوفيق البيانات المشاهدة سيكون غير ملائم. والسبب في ذلك هو أن أي خط سوف يتجاوز بالضرورة الواحد الصحيح ويسقط دون الصفر أيضاً إلاّ إذا كان الميل يساوي صفر.

شكل (3): خط توفيق العلاقة الخطية بين متغيّر متصل وآخر ثنائي القيمة



تحويلات الانحدار اللوجستي

هناك عدّة إجراءات تحويلية يمكن أن تقدّم مساهمات جدّية لحل بعض الصعوبات والتحديات التي ذكرت سابقاً، وسيقوم الباحث بعرض أهم المفاهيم التي ستسهم في تقديم تلك الحلول على النحو التالي:

الاحتمال Probability

الاحتمال هو عبارة عن عدد يتراوح ما بين الصفر والواحد الصحيح، وهي تعبر عن أرجحية likelihood وقوع حدث معين. فعلى سبيل المثال، احتمال فوز فريق ما هو عبارة عن عدد مرات الفوز مقسوماً على العدد الكلي للمباريات، وبهذا المعنى فإن الاحتمال هو عبارة عن حاصل قسمة عدد مرات النجاح على العدد الكلي للمحاولات (Walker,1996,P.34). وحيث إنّ المتغيّر التابع في حالة هذه الدراسة هو متغيّر ثنائي القيمة يأخذ إحدى القيمتين $Y=1$ لظهور السمة و $Y=0$ عند عدم ظهورها، فإنّنا نلاحظ أنّه إذا جمعنا جميع الحالات التي تكون فيها $Y=1$ وقسمناها على العدد الكلي للحالات، فإنّ القيمة الناتجة تمثّل متوسط قيمة المتغيّر التابع Y، وهذه القيمة تقابل تماماً نسبة أو احتمال أن تكون قيمة المتغيّر التابع يساوي واحداً $Y=1$ في بيانات العينة.

وبناء على ذلك فإنّ الخطوة الأولى لتوفيق البيانات بين المتغيّر أو المتغيّرات المستقلة والمتغيّر التابع الثنائي Y، هو التعامل مع المتغيّر التابع ثنائي القيمة كما لو كان متغيّراً متصلاً بحيث إنّ القيم المتوقعة له تمثّل احتمال أن يكون المتغيّر التابع يأخذ القيمة $Y=1$ ، وليس المتغيّر التابع نفسه والذي

لا يأخذ إلاّ إحدى القيمتين صفر أو واحد. إنّ توفيق بيانات المتغيّرات المستقلة X's مع احتمال أن يكون المتغيّر التابع يأخذ القيمة واحداً P(Y=1) بدلاً من المتغير التابع Y نفسه يفتح الطريق للتعامل مع المتغيّر التابع P(Y=1) كمتغيّر متصل.

إنّ الطريقة السابقة التي تتمثّل في اعتبار المتغيّر التابع هو احتمال أن يمتلك صفة ما تم ترميزها بالقيمة واحد أي P(Y=1) بدلاً من المتغيّر التابع ذاته Y ، وكذلك فتح الطريق للتعامل مع المتغيّر التابع المعدّل كمتغيّر متصل بدلاً من كونه في الأصل متغيّراً ثنائي القيمة بحيث يمكن توفيقه بنموذج خطي، كل ذلك قد أوجد مشكلة مفاهيمية خطيرة، وهي إمكانية ظهور قيم متوقعة للمتغيّر التابع الجديد تتجاوز الواحد صحيح أو تقل عن الصفر، وهو ما يتناقض مع مفهوم الاحتمالات، الأمر الذي يجعل من الخطأ بناء معادلة خطية للتنبؤ بالاحتمال P(Y=1) (Dallal, 2001; Cizek & Fitzgeral,1999).

إنّ إحدى طرق التعامل مع هذه المشكلة والمتمثّلة في كون متغيّر الاستجابة مقيّداً بقيم محدّدة (من صفر إلى واحد صحيح) هي تطوير دالة استجابة محوّلة تستطيع أخذ أي قيمة، وتستخدم التوليفة الخطية للمتغيرات المستقلة، ولذا فإنّ الخطوة الأولى لتحقيق ذلك هو إجراء تحويل بسيط ومهم يتمثّل في استخدام معامل الترجيح Odds بدلاً من الاحتمالات P .

معامل الترجيح **Odds**

إنّ معامل الترجيح Odds هو عبارة عن طريقة للتعبير عن مدى احتمال حدوث شيء ما مقارنة باحتمال عدم حدوثه، وغالباً ما يتم التعبير عنه على شكل نسبة بين العددين. فإذا توقع شخص فوز فريق ما في ثلاث من خمس مباريات، وفوز الفريق الآخر في مباراتين من المباريات الخمس، فهذا يعني أنّ احتمال فوز الفريق الأول هو $\frac{3}{5} = 0.60$ واحتمال فوز الفريق الثاني هو $\frac{2}{5} = 0.40$.

تحويل معامل الترجيح **Odds** إلى دالة اللوجت **Logit** :

إذا تم أخذ اللوغاريتم الطبيعي لمعامل الترجيح O نلاحظ ما يأتي:

$$\because O = \frac{P}{1-P} = e^{a+b_1 x_1}$$

$$\therefore \ln \ Odds = \ln\left(\frac{P}{1-P}\right) = a + b_1 \ x_1 \qquad (17)$$

حيث: ln Odds هو لوغاريتم معامل الترجيح.

a هي معامل الثابت، و b_1 هي معامل المتغيّر المستقل X_1.

$$\therefore -\infty < \ln\left(\frac{P}{1-P}\right) < +\infty$$

لاحظ أن أخذ اللوغاريتم الطبيعي لمعامل الترجيح O جعل العلاقة بين المتغير التابع (والذي هو في هذه الحالة (ln(Odds) والمتغير المستقل X_1 علاقة خطية تتمتع بخاصة الإضافة additive. كما يلاحظ أن الحد الأدنى للقيم المسموح بها لمعامل الترجيح والتي كانت تساوي صفراً، أصبح يقابلها في لوغاريتم معامل الترجيح (ln(Odds القيمة سالب ما لانهاية (∞ -). وهذا يعني عندما تكون قيمة معامل الترجيح الواحد الصحيح، فإن قيمة لوغاريتم

معامل الترجيح المقابل له هي صفر، أما إذا كان معامل الترجيح أكبر من الواحد الصحيح، فإن قيمة لوغاريتم معامل الترجيح المقابل له هي عدد موجب، أما إذا كان معامل الترجيح يساوي أقل من الواحد الصحيح، فإن لوغاريتم معامل الترجيح المقابل له سوف يكون عدداً سالباً وهكذا.

تفسير معاملات الانحدار اللوجستي

يرى Pample(2000) بأنّه حسب المتوقع والمعتاد من التحويلات غير الخطية، فإن تأثيرات المتغيرات المستقلة على المتغير التابع في تحليل الانحدار اللوجستي ستكون لها عدة تفسيرات، وأنّ تأثيرات المتغيرات المستقلة ستكون حاضرة على الاحتمالات، ومعاملات الترجيح، ولوغاريتمات معاملات الترجيح، وأنّ التفسير بناء على أيّ ممّا سبق له إيجابياته وسلبياته

(أ) تفسير المعاملات بدلالة اللوجت

وهي طريقة مباشرة للتفسير باستخدام معاملات الانحدار اللوجستي التي تم تقديرها. فمعاملات الانحدار اللوجستي توضح ببساطة التغير في لوغاريتمات معاملات الترجيح المتوقعة لكل تغير بمقدار وحدة واحدة في المتغيرات المستقلة (Dallal,2001). وبذلك فإنه في هذه الطريقة يكون للمعاملات تفسيراً مطابقاً لما هو عليه الأمر في تحليل الانحدار الخطي، والفرق الوحيد هو في وحدات المتغير التابع، حيث إن وحدات المتغير التابع في هذه الحالة تمثّل لوغاريتمات معاملات الترجيح (Cizek & Fitzgerald,1999). أي أنّ معاملات الانحدار في كلتا الحالتين تمثّل العلاقة

بـين المتغيّـر المسـتقل أو المتغيّـرات المسـتقلة والمتغيّـر التـابع ملخّصـة بقيمـة إحصائية واحدة هـي قيمـة المعامـل، وذلـك بغض النظـر عـن مسـتويات قـيم المتغيّـر أو المتغيّـرات المستقلة. أي أنّـه إذا كـان لـدينا متغيّـر مسـتقل واحـد في النمـوذج هـو X، فـإنّ التغيّـر بمقـدار وحـدة واحـدة مـن ذلـك المتغيّـر المسـتقل سيكون لـه نفس التأثير في المتغيّر التابع Y سواء كنّا نتحدّث عن قيم عالية أو متوسطة أو منخفضة للمتغيّر المستقل X.

(ب) تفسير المعاملات بدلالة معاملات الترجيح

وهـي طريقـة لتفسـير معـاملات الانحـدار اللوجسـتي تنبـع مـن تحـويلات النمـاذج اللوجسـتية، بحيـث إن المـتغيرات المسـتقلة تـؤثر علـى معامـل الترجيـح

بدلاً من تأثيرها على لوغاريتم معامل الترجيح للمتغير التابع. وللحصول على تـأثيرات المـتغيرات المسـتقلة علـى معـاملات الترجـيح، تؤخـذ الدالـة الأسـية exponent(e) للوجت أي معكوس لوغاريتم معاملات الترجيح.

فعلى سبيل المثال. في حالة النموذج البسيط، إذا تم أخذ الدالـة الأسية للطـرفين exponent(e) ، فـإن ذلـك يزيـل اللوغـاريتم عـن معـاملات الترجـيح، وبذلك يظهر أثر المتغيرات المستقلة على معامل الترجيح.

$$\because \ln\left(\frac{P}{1-P}\right) = b_0 + b_1 x_1 + b_2 x_2$$

$$\therefore e^{\ln\left(\frac{P}{1-P}\right)} = e^{b_0 + b_1 x_1 + b_2 x_2}$$

$$\therefore Odds = \frac{P}{1-P} = e^{b_0 + b_1 X_1 + b_2 X_2}$$

توضح المعادلة السابقة العلاقة بين X's ومعامل الترجيح. وكما هو واضح، فإن معكوس اللوغاريتم للوغاريتم يساوي المقدار نفسه (أي عامل الترجيح) كما هو في الطرف الأيسر من المعادلة السابقة. وبما أن $e^{(x+y)}$ تساوي $e^{x} * e^{y}$ كما هو في الطرف الثاني من المعادلة السابقة، فإن المعادلة أصبحت خاضعة لخاصية الضرب Multiplicative بدلاً من خاصية التجميع additive . إن معاملات الترجيح هي دالة لـ e^{b_0} و $e^{b_1 x_1}$ و $e^{b_2 x_2}$ ، أي أنّ تأثير كل متغيّر مستقل على معامل الترجيح يعرف من خلال أخذ معكوس لوغاريتم المعاملات. وببساطة، فإنّ معاملات الترجيح Odds هي دالة لـ e^{b_0} مضروبة في $e^{b_1 x_1}$ مضروبة في $e^{b_2 x_2}$ ،وهكذا حسب عدد المتغيّرات المستقلة في النموذج. ومع أن أكثر برامج التحليل الإحصائي الحاسوبية لا تعرض هذا النوع من الإجراءات في مخرجاتها، إلا أن المهتم يمكنه الحصول على هذه النتائج باستخدام الحاسبات الآلية، وذلك بحساب e^{x} (Pample,2000,P.21).

التعقيد هو أن تأثير العوامل المختلفة على معامل الترجيح أصبح خاضعاً لخاصية الضرب بدلاً من خاصية الجمع. ففي معادلات الانحدار الخطي الاعتيادي والتي تخضع لخاصية الجمع كما في المعادلة التالية: $y = b_0 + b_1 x_1 + b_2 x_2$ ، فإن المتغير الذي تكون قيمة معامل انحداره تساوي صفراً لن يؤثر في المتغير التابع، وذلك لأن حاصل ضرب المتغير X بالمعامل الذي قيمته تساوي صفراً سينتج مقداراً قيمته تساوي الصفر. وعند جمع هذا الحد مع حاصل ضرب بقية العوامل في متغيراتها، نجد أن تأثير ذلك الحد سيكون معدوماً لأنه صفر، ولن يؤثر في القيمة المتوقعة لـ Y.

الفصل الثامن

تحليـــل البقـــاء

Survival Analysis

- مشاكل التحليل البقائي
- آليات الاختفاء
- أنواع الاختفاء
  ١. اختفاء من النوع الأول
  ٢. اختفاء من النوع الثاني
  ٣. الاختفاء العشوائي
- مصطلحات ورموز
- دوال البقاء الأساسية

تحليل البقاء Survival Analysis

نماذج تحليل البقاء تتناول الزمن الذي يسبق حدوث حدث معين، ومن أكثر الأمثلة تطبيقا في هذا المجال هو الزمن الذي يسبق الوفاة، ولكن تطبيق تحليل البقاء هو أوسع من ذلك بكثير، فهو يستخدم في كثير من المجالات والتخصصات المختلفة والتي تعتبر الزمن عامل أساسي في تحليل الظاهرة المعنية بالدراسة. والميزة الأساسية في هذا الأسلوب هو دراسة العلاقة بين الزمن الذي يسبق حدوث الحدث مع متغير أو أكثر من المتغيرات المستقلة بغض النظر عن طبيعة هذه المتغيرات من حيث كونها كمية أو وصفية أو مختلطة. (Fox, 2002).

في البداية نبدأ بوصف نوع المشكلة التي يتناولها تحليل البقاء .

بشكل عام تحليل البقاء هو عبارة عن مجموعة من الإجراءات الإحصائية لتحليل بيانات يكون المتغير محل الاهتمام هو الزمن الذي يمر حتى حدوث الحدث .

نعني بالزمن Time : سنوات، أشهر، أيام، أسابيع ... ، يبدأ هذا الزمن بداية فترة المتابعة حتى حدوث الحدث .

أو الزمن هو عمر المفردة عند حدوث الحدث .

أما الحدث Event : فنحن نقصد بالحدث أي حث يحدث للمفردة سواء ايجابي أو سلبي فقد يعبر عن الحدث بالموت مثلا ، أو حدوث المرض، العودة إلى العمل، الإنعاش ...أي حدث يحدث للمفردة .

لاحظ أنه قد يحدث أكثر من حدث للمفردة أثناء أو في نفس التحليل ، لكننا نفترض أن حدث واحد فقط هو محل الاهتمام .

مشاكل التحليل البقائي

(أمثلة على المواضيع التي يتناولها تحيل البقاء)

- دراسة مرضى سرطان الدم خلال عدة أسابيع لمعرفة ما هي مدة بقاءهم في المرض حتى حدوث الموت ، فهنا يكون المتغير التابع هو الزمن الذي يمر حتى حدوث الحدث ، فالحدث هنا يعبر عنه بالموت .
- دراسة مجموعة من الأفراد الخالين من الأمراض خلال عدة سنوات لمعرفة كيف تطورت أمراض القلب ، (الزمن الذي يمر حتى الإصابة بمرض القلب) .
- اعتبار 13 سنه متابعة للأشخاص المسنين (60 سنه فأكثر) لمعرفة ما هي أسباب طول مدة بقائهم على قيد الحياة (الزمن الذي يمر بالسنوات حتى حدوث الوفاة).
- المساجين المفرج عنهم حديثا وبشروط ولعدة أسابيع لمعرفة ما إذا كان قد أعيد أعقاله (الزمن الذي يمر بالأسابيع حتى إعادة اعتقاله) .
- ما هي مدة بقاء المرضى بعد تلقيهم عملية زراعة القلب (الزمن الذي يمر بالأشهر حتى حدوث الوفاة).

جميع الأمثلة السابقة هي مشاكل تحليل البقاء لأن المتغير الناتج هو الزمن الذي يمر حتى حدوث الحدث .

في تحليل البقاء نحن بالغالب نشير إلي متغير الزمن بزمن البقاء ، لأنه يعطي الزمن للمفردات التي بقيت خلال فترة المتابعة .

ويقول (Rodriguez, 2001) أن تحليل البقاء يتعلق بتحليل البيانات التي تحتوي على ثلاث خصائص رئيسية وهي:

١- المتغير التابع هو زمن البقاء حتى حدوث الحدث.

٢- وجود بيانات اختفاء Censored Data .

٣- وجود متغيرات مفسرة يُفترض أنها تؤثر على زمن البقاء .

بيانات زمن البقاء تحتوي على جزء رئيسي ومميز وهو الاختفاء والذي قد يسبب قلق وإزعاج في التحليل إذا لم يكن مسيطر عليه بشكل كاف. ووجود بعض المشاهدات المختفية في بيانات البقاء لا يمكن تجاهلها أو إهمالها (Filler, 2003).

آليات الاختفاء Censoring Mechanisms

نحن نقصد بالاختفاء هو وجود مفردات لا نعرف زمن حدوث الحدث لها ولا نستطيع تتبعها خلال فترة زمنية ، والاختفاء يتكرر كثيراً في بيانات البقاء فهناك بعض المفردات يحدث لها الحدث وبالتالي يمكننا تحديد زمن البقاء لها والبعض الآخر ليس لدينا معلومات كافية عنه، وشرط استخدام الاختفاء هو أن يكون الاختفاء مستقل ولا يعتمد على خطر التجربة.

هناك ثلاثة أسباب لحدوث الاختفاء :

١. لا يحدث الحدث قبل انتهاء الدراسة .

٢. المفردة فقدت أثناء المتابعة خلال فترة الدراسة.

٣. انسحاب أو خروج المفردة من الدراسة ، بسبب الوفاة (إذا لم يكن الموت من مصلحة الحدث المدروس) أو لأي سبب أخر.

أنواع الاختفاء

يوجد هناك أنواع مختلفة من الاختفاء ولكن أهم الأنواع الشائعة هي:

أولاً : اختفاء من النوع الأول Type I Censoring

في هذا النوع تكون المدة الكلية للدراسة ثابتة بينما عدد الأحداث (أي عدد الأفراد الذين حدث لهم الحدث) يكون متغير عشوائي، ويدعى هذا النوع من الاختفاء بـ (الاختفاء الثابت) وفيه يتم تحديد زمن إيقاف الدراسة بعد فترة زمنية محددة .

والاختفاء من النوع الأول يكون واحد من الأنواع التالية:

- **الاختفاء الأيمن** Right Censoring (Suspended)

وهي الحالة الأكثر شيوعاً في بيانات البقاء، وهذه الحالة تكون مرتبطة بالمفردات التي لم يحدث لها الحدث. بعض المفردات تبقى على قيد الحياة عند نهاية الدراسة أي أن زمن البقاء للمفردة يفوق نقطة انتهاء الدراسة وهذه المفردات يقال عنها اختفاء أيمن.

أسباب حدوث الاختفاء الأيمن:

١- قرار الباحث إنهاء الدراسة قبل حدوث الحدث.

٢- عدم القدرة على الوصول للمفردة لأي سبب.

٣- بعض المفردات لم يحصل لها الحدث.

- **الاختفاء الأيسر** Left Censoring

لنفرض أن لدينا مفردة دخلت في الدراسة لكن زمن التعرض للخطر غير معلوم بينما زمن حدوث الحدث هو المعلوم فقط ، ومثال على ذلك مرضى السرطان ومرضى الايدز فان زمن بداية الإصابة بالمرض غير معلوم لكن زمن الوفاة بسبب ذلك المرض هو المعلوم، وهذه المفردة يقال لها اختفاء أيسر.

- **الاختفاء الفتري** Interval Censoring

وفي هذه الحالة يكون زمن حدوث الحدث بالضبط غير معلوم لبعض المفردات لكن المعلوم هو الفترة الزمنية التي وقع فيها الحدث، ويقال عن هذه المفردات اختفاء فتري.

ثانياً : اختفاء من النوع الثاني Type II Censoring

في هذه الحالة يكون عدد المفردات التي يحدث لها الحدث معروف (ثابت) مقدماً بينما فترة الدراسة الكلية تكون متغير عشوائي لا يمكن معرفتها مقدماً. وفيها يتم تحديد زمن انتهاء الدراسة بعد عدد معين من حالات حدوث الحدث.

ثالثاً : الاختفاء العشوائي Random Censoring (Hybrid)

في هذا النوع كل مفردة لها زمن اختفاء متوقع Ci وعمر (زمن بقاء) متوقع Ti ويفترض أن زمن الاختفاء وزمن البقاء متغيرين عشوائيين مستقلين، ونلاحظ أن Yi = min: (Ci, Ti) أي هو زمن البقاء أو زمن الاختفاء أيهما أقل ، وأن المتغير المؤشر يدعى di ويخبرنا بأن المشاهدة انتهت بالوفاة أو بالاختفاء وهذا النوع يعتبر مزيج من النوعين السابقين.

المصطلحات والرموز

سنتناول المصطلحات والرموز الأساسية لتحليل البقاء

T هي متغير عشوائي لزمن بقاء المفردات ، وتأخذ القيم الموجبة فقط ،لأن الزمن بالموجب وليس هناك زمن بالسالب. $T >= 0$.

بينما تشير t إلي قيمة محددة للمتغير العشوائي T .

مثال : إذا كنا مهتمون بتقييم ما إذا كان شخص على قيد الحياة لأكثر من 5 سنوات وذلك بعد خضوعه لعلاج السرطان فإن $t = 5$ ويمكن أن نسأل هل $T > 5$.

ونستخدم هنا الحرف اليوناني دلتا δ ويشير إلى $(1 \; , \; 0)$ وهو متغير عشوائي يدل على الفشل (الحدث) أو الاختفاء ، حيث $\delta = 1$ تشير إلي الحدث إذا ظهر خلال فترة المتابعة ، و $\delta = 0$ إذا زمن البقاء يكون اختفاء ، وذلك لانتهاء فترة الدراسة أو لفقدان المفردة أو لانسحابها .

$$\delta = (0, 1) = \begin{cases} 1 & \text{if failure} \\ 0 & \text{if censored} \end{cases}$$

- study ends
- lost to follow-up
- withdraws

دوال البقاء الأساسية Basic Survival Functions

لنفرض أن T هي متغير عشوائي متصل موجب ويمثل زمن البقاء حتى حدوث الحدث وله دالة كثافة احتمال $f(t)$ ودالة توزيع تراكمية $F(t)$، وبذلك تكون دوال البقاء الأساسية:

١- دالة البقاء: The Survival Function

وتعرف أيضاً بدالة الصلاحية The Reliability function ويشار إليها بـ$S(t)$ وتعرف دالة البقاء على أنها احتمال البقاء إلى ما بعد t، (على الأقل حتى الزمن t)

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx$$

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}(1 - S(t)) = -S'(t)$$

$$\frac{dS(t)}{dt} = -f(t) \quad [= -\frac{dS(t)}{dt}]$$

٢- دالة الخطر: The Hazard Function

وتعرف بمعدل الخطر Hazard Rate أو المعدل اللحظي أو الحالي لظهور الحدث ويشار إليها بـ $h(t)$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}$$

ونرى بأن البسط لهذا الكسر السابق هو الاحتمال الشرطي أن الحدث سيظهر في الفترة بين $(t, \; t+t\Delta)$ بشرط عدم ظهوره قبل t، والمقام هو طول الفترة .

٣- دالة الخطر التراكمية: The Cumulative Hazard Function

وتعرف دالة الخطر التراكمية على أنها مجموع الأخطار التي حدثت حتى الزمن t.

$$= \int_0^t h(x)\,dx \quad H(t)$$

$$= \int_0^t \frac{f(x)}{S(x)}\,dx$$

$$= - \int_0^t \frac{1}{S(x)} \left\{ \frac{d}{dx} S(x) \right\} dx$$

$$(S(t)) = -\text{Ln}$$

وبناءاً على ما سبق فإن:

$$S(t) = \exp(-H(t))$$

$$F(t) = 1 - \exp(-H(t))$$

$$f(t) = h(t).\exp(-H(t))$$

٤- توقع الحياة: The Expectation of Life

لنفرض أن μ تشير إلى متوسط أو توقع القيمة T حيث أن:

$$\mu = \int_0^\infty t \,.\, f(t)\,dt$$

وباستخدام عملية التكامل بالتجزئ نتوصل إلى:

$$\mu = \int_0^\infty S(t)\,dt \qquad ; \quad S(0) = 1 \ \& \ S(\infty) = 0$$

حيث أن S(t) تعطي الاحتمال بأن المفردة تبقى على قيد الحياة بعد t، بينما E(t) هي توقع الحياة للمفردة.

بالنسبة للدالتين اللتان تم دراستهما $S(t)$ و $h(t)$ فإن :

$S(t)$ دالة البقاء أكثر طبيعية وجاذبية لتحليل بيانات البقاء، وذلك لسبب بسيط لان $S(t)$ تصف مباشرة تجربة البقاء لفوج الدراسة .

وبالنسبة لدالة الخطر $h(t)$ فهي أيضا مهمة للأسباب التالية :

- **دالة الخطر هي قياس المعدل اللحظي ، بينما منحنى البقاء هو قياس تراكمي على مر الزمن.**
- **دالة الخطر يمكن استخدامها لتحديد شكل النموذج ، مل المنحنى الأسي أو منحنى ويبل أو اللوغرتمي الطبيعي ...**
- **دالة الخطر هي وسيلة للنماذج الرياضية لبيانات البقاء التي أجريت(نفذت) ، وعادة ما يكتب نموذج البقاء بشروط دالة الخطر.**